

Speech Rhythm

The language-
specific integration
of pitch and
duration

Ruth E. Cumming

Downing College

This thesis is submitted for the degree
of Doctor of Philosophy

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except where specifically indicated in the text.

79,130 words

Summary

Experimental phonetic research on speech rhythm seems to have reached an impasse. Recently, this research field has tended to investigate produced (rather than perceived) rhythm, focussing on timing, i.e. duration as an acoustic cue, and has not considered that rhythm perception might be influenced by native language. Yet evidence from other areas of phonetics, and other disciplines, suggests that an investigation of rhythm is needed which (i) focuses on listeners' perception, (ii) acknowledges the role of several acoustic cues, and (iii) explores whether the relative significance of these cues differs between languages. This thesis, the originality of which derives from its adoption of these three perspectives combined, indicates new directions for progress.

A series of perceptual experiments investigated the interaction of duration and f₀ as perceptual cues to prosody in languages with different prosodic structures – Swiss German, Swiss French, and French (i.e. from France). The first experiment demonstrated that a dynamic f₀ increases perceived syllable duration in contextually isolated pairs of monosyllables, for all three language groups. The second experiment found that dynamic f₀ and increased duration interact as cues to rhythmic groups in series of monosyllabic digits and letters; the two cues were significantly more effective than one when heard simultaneously, but significantly less effective than one when heard in conflicting positions around the rhythmic-group boundary location, and native language influenced whether f₀ or duration was the more effective cue.

These two experiments laid the basis for the third, which directly addressed rhythm. Listeners were asked to judge the rhythmicity of sentences with systematic duration and f₀ manipulations; the results provide evidence that duration and f₀ are interdependent cues in rhythm perception, and that the weighting of each cue varies in different languages. A fourth experiment applied the perceptual results to production data, to develop a rhythm metric which captures the multi-dimensional and language-specific nature of perceived rhythm in speech production. These findings have the important implication that if future phonetic research on rhythm follows these new perspectives, it may circumvent the impasse and advance our knowledge and model of speech rhythm.

Table of contents

Chapter 1	Introduction.....	1
1.1	What is rhythm?.....	1
1.2	Review of speech-rhythm research.....	2
1.2.1	Pre-1900.....	3
1.2.2	1900-1939.....	3
1.2.2.1	Psychology.....	3
1.2.2.2	Phonetics.....	4
1.2.2.3	Summary (1900-1939).....	4
1.2.3	1940s-1970s.....	5
1.2.3.1	Rhythm typology.....	5
1.2.3.2	Isochrony investigation.....	6
1.2.3.3	Non-isochrony-based research.....	8
1.2.3.4	Summary (1940s-1970s).....	8
1.2.4	1980s-1990s.....	9
1.2.4.1	Stress-timing/syllable-timing rejected.....	9
1.2.4.2	Alternative rhythm typologies: categorical or continuous.....	9
1.2.4.3	Perceptual isochrony and P-centres.....	11
1.2.4.4	Phonological models.....	12
1.2.4.5	Infant studies.....	13
1.2.4.6	Summary (1980s-1990s).....	14
1.2.5	Twenty-first century.....	14
1.2.5.1	Widespread use of rhythm metrics.....	14
1.2.5.2	Other phonetic research.....	18
1.2.5.3	Other disciplines.....	19
1.2.5.4	Summary (twenty-first century).....	21
1.2.6	Current state of experimental phonetic research on rhythm.....	21
1.2.6.1	Empirical investigation of perception.....	23
1.2.6.2	Measurement of acoustic cues other than duration.....	23
1.2.6.3	Investigation of native-language influence.....	24
1.3	Aims and research questions.....	27
1.4	Outline of thesis.....	28
Chapter 2	The languages investigated: Swiss German; Swiss French; French.....	29
2.1	Why Switzerland?.....	29
2.2	Social context.....	30
2.2.1	Swiss German.....	31
2.2.2	(Swiss) French.....	32
2.2.3	Summary (social context).....	34
2.3	Prosody.....	34
2.3.1	Swiss German.....	35
2.3.1.1	Rhythm.....	35
2.3.1.2	Prominence.....	36
2.3.1.2.1	Phonology.....	36

2.3.1.2.2	Phonetic correlates.....	36
2.3.1.3	Intonation.....	37
2.3.1.4	Cross-dialectal variation.....	39
2.3.2	French.....	40
2.3.2.1	Rhythm.....	40
2.3.2.2	Prominence.....	41
2.3.2.2.1	Phonology.....	41
2.3.2.2.2	Phonetic correlates.....	42
2.3.2.2.3	Controversy over French prominence.....	43
2.3.2.3	Intonation.....	45
2.3.2.4	Swiss French.....	47
2.3.3	Summary (prosody).....	49
Chapter 3	The influence of dynamic f0 on the perception of duration.....	51
3.1	Summary.....	51
3.2	Previous research.....	51
3.3	Extending previous research.....	52
3.3.1	Listeners' native language.....	52
3.3.2	Experiment design (previous studies).....	53
3.3.2.1	Listeners' task.....	53
3.3.2.2	Linguistic versus non-linguistic stimuli.....	54
3.3.2.3	Dynamic f0 in stimuli: direction, excursion, timing.....	55
3.3.2.4	Duration of stimuli.....	56
3.4	Hypothesis.....	56
3.4.1	Swiss German.....	56
3.4.2	(Swiss) French.....	57
3.4.3	Alternatives.....	57
3.5	Method.....	57
3.5.1	Subjects.....	58
3.5.2	Stimuli.....	58
3.5.3	Trials.....	60
3.5.4	Procedure.....	61
3.5.5	Analysis.....	62
3.6	Results.....	62
3.6.1	Level stimuli: fillers and controls.....	62
3.6.2	Dynamic stimuli.....	64
3.6.3	Further analysis.....	65
3.6.3.1	Direction of f0 movement.....	67
3.6.3.2	Timing of f0 movement.....	69
3.6.3.3	Duration of stimuli.....	71
3.6.3.4	Order of stimuli.....	72
3.6.3.5	f0 height.....	74
3.7	Discussion.....	76
3.7.1	Individual languages.....	77
3.7.2	Human language.....	78

3.7.3	Nature of stimuli.....	79
3.8	Conclusion.....	81
Chapter 4	The interdependence of f0 and duration in rhythmic-group perception	83
4.1	Summary.....	83
4.2	Prosodic groups.....	83
4.3	Hypothesis.....	85
4.4	Subjects.....	86
4.5	Method.....	89
4.5.1	Recordings.....	89
4.5.2	Stimulus preparation.....	92
4.5.2.1	Token selection.....	93
4.5.2.2	'Base' syllables.....	94
4.5.2.3	Duration/f0 manipulations.....	94
4.5.2.4	Concatenation of syllables.....	96
4.5.3	Procedure.....	97
4.5.4	Analysis.....	98
4.6	Results.....	100
4.6.1	Control stimuli.....	100
4.6.2	Initial analysis: all variables.....	102
4.6.3	Digit/letter pattern.....	104
4.6.4	Main analysis: cue(s); native language(s).....	106
4.6.4.1	Monolinguals.....	107
4.6.4.2	Bilinguals.....	108
4.6.4.3	'Conflicting' stimuli.....	110
4.7	Discussion.....	111
4.7.1	Interdependence of duration and f0.....	111
4.7.2	Native language.....	113
4.7.3	Other findings.....	114
4.8	Conclusion.....	115
Chapter 5	The interdependence of f0 and duration as cues to the perceived rhythmicity of sentences.....	117
5.1	Summary.....	117
5.2	Previous research.....	117
5.2.1	How to test rhythm perception.....	117
5.2.2	Evidence for interdependence of duration and f0.....	120
5.3	Pilot experiment.....	121
5.4	Main experiment: stimulus design.....	122
5.4.1	Structure of sentences.....	122
5.4.2	Duration and f0 values chosen for manipulations.....	125
5.4.2.1	AXB discrimination pre-test.....	126
5.4.3	Preparation of stimuli.....	127
5.5	Hypothesis.....	130
5.6	Main experiment (A): preliminary test.....	133

5.6.1	Procedure.....	133
5.6.2	Preliminary test: results.....	134
5.7	Main experiment (B): final version.....	136
5.7.1	Subjects.....	136
5.7.2	Procedure.....	137
5.7.3	Results.....	137
5.7.3.1	Rhythmicality judgements.....	137
5.7.3.2	Intra- and inter-subject consistency.....	143
5.7.3.3	Post-test questionnaire responses.....	148
5.8	Discussion.....	152
5.8.1	Interdependence of duration and f_0	152
5.8.2	Influence of native-language (prosodic) phonology on perceived rhythm	152
5.8.3	Cross-linguistic differences in stimuli.....	155
5.8.4	Other factors linked to inter-subject variation.....	156
5.9	Conclusion.....	158
Chapter 6	PVIs which account for the acoustic multi-dimensionality and language-specificity of perceived rhythm.....	160
6.1	Summary.....	160
6.2	Rhythm metrics.....	160
6.2.1	Problems.....	160
6.2.2	Possible solution.....	161
6.3	Hypothesis.....	162
6.3.1	Durational PVIs.....	162
6.3.2	Tonal PVIs.....	163
6.3.3	Weighted PVIs.....	163
6.4	Method.....	164
6.4.1	Reading text.....	164
6.4.2	Subjects and procedure.....	164
6.4.3	Analysis.....	165
6.4.3.1	Acoustic measurements.....	165
6.4.3.2	Weighting values.....	168
6.4.3.3	PVI calculations.....	172
6.5	Results.....	176
6.5.1	PVIs.....	176
6.5.2	Other analyses.....	181
6.6	Discussion.....	183
6.6.1	(Vowel or syllable) durational PVIs.....	183
6.6.2	Tonal PVIs.....	185
6.6.3	Weighted PVIs.....	186
6.6.4	Inter-speaker variation.....	187
6.7	Conclusion.....	189
Chapter 7	Conclusions.....	191
7.1	Contributions to speech-rhythm research.....	191

7.1.1	A focus on perception.....	192
7.1.2	Inclusion of f0.....	194
7.1.3	Cross-linguistic study.....	195
7.2	Future research.....	197
7.3	Consequences of this research.....	199
7.3.1	Implications for theoretical models.....	199
7.3.1.1	Modelling cross-linguistic variation in rhythm.....	199
7.3.1.2	Speech perception models.....	202
7.3.2	Implications for practical applications.....	204
7.3.2.1	Speech technology.....	204
7.3.2.2	L2 teaching.....	205
7.3.2.3	Remediation of disorders in L1 acquisition.....	207
7.4	Does rhythm exist in speech?.....	208
Appendices.....		211
8.1	Experiment 1 (chapter 3).....	212
8.1.1	Instructions.....	212
8.1.2	Results.....	213
8.2	Experiment 2 (chapter 4).....	215
8.2.1	Instructions.....	215
8.2.2	Results.....	216
8.3	Experiment 3 (chapter 5).....	217
8.3.1	Stimulus sentences.....	217
8.3.2	Instructions.....	219
8.3.3	Post-test questionnaire.....	220
8.3.4	Results.....	221
8.4	Experiment 4 (chapter 6).....	222
8.4.1	Texts.....	222
8.4.2	Results.....	223
Bibliography.....		224

Table of figures

Figure 2-1	Map of Switzerland showing data from the federal census in 2000.....	30
Figure 2-2	Comparison of standard German and SstG intonation patterns.....	38
Figure 3-1	Percentage of ‘1 st sound is longer than 2 nd sound’ responses compared to the level of chance.....	63
Figure 3-2	Percentage of ‘dynamic is longer than level’ (‘D>L’) responses compared to the level of chance.....	64
Figure 3-3	Percentage (out of total number possible) of ‘D>L’ responses depending on direction of f0 change.....	68
Figure 3-4	Percentage (out of total number possible) of ‘D>L’ responses depending on timing and direction of f0 change.....	70
Figure 3-5	Percentage (out of total number possible) of ‘D>L’ responses depending on duration of stimuli.....	71
Figure 3-6	Percentage (out of total number possible) of ‘D>L’ responses depending on whether the dynamic stimulus was heard first or second.....	73
Figure 3-7	Percentage of ‘high>low’ responses compared to the level of chance.....	75
Figure 4-1	For each syllable (1-5): mean duration across 14 sequences.....	91
Figure 4-2	For each syllable (1-5): across 14 sequences, mean f0 excursion; minimum/maximum f0 excursion; mean average f0.....	92
Figure 4-3	Six manipulations and the ‘base’ form.....	95
Figure 4-4	Stimuli: sequences of durationally/tonally manipulated syllables spliced together.....	96
Figure 4-5	Mean percentage (across subjects) of 3+2 responses to control stimuli.....	100
Figure 4-6	Mosaic diagram: areas of rectangles represent the DS (summed over subjects) for each level of three variables (language(s), cue(s), pattern).....	102
Figure 4-7	Mean percentage DS across all subjects per language group for each cue condition and digit/letter pattern.....	106
Figure 4-8	Percentage of responses to ‘conflicting’ stimuli in which subjects used either length or pitch.....	110
Figure 4-9	Percentage of subjects who used length in 0-3, 4-6, 7-9 and 10-12 of the 12 ‘conflicting’ stimuli for each pitch condition.....	111
Figure 5-1	f0 contours for pilot stimuli.....	121
Figure 5-2	Example sentence displayed as spectrogram and waveform.....	128
Figure 5-3	Sentence in Figure 5-2 converted to ‘manipulation object’ with waveform, glottal pulses, <i>PitchTier</i> and <i>DurationTier</i>	128
Figure 5-4	<i>PitchTiers</i> showing the peak point moved for the F0 _{High} and F0 _{Low} stimuli.....	129
Figure 5-5	<i>DurationTiers</i> with section specified to be shortened/lengthened.....	130
Figure 5-6	Hypothetical data: if F0 _{Norm} /DUR _{Norm} stimulus were preferred for all 30 sentences.....	131
Figure 5-7	Hypothetical data: if subjects prefer non-deviant f0, but are more tolerant of durational deviance.....	132
Figure 5-8	Hypothetical data: if subjects prefer non-deviant duration, but are more tolerant of tonal deviance.....	132
Figure 5-9	Preliminary test: mean (across subjects) frequency of responses per stimulus type.....	135
Figure 5-10	Mean (across subjects) frequency of responses per stimulus type (1).....	138
Figure 5-11	Mean (across subjects) frequency of responses per stimulus type (2).....	140
Figure 5-12	Mean (across subjects) frequency of responses per stimulus type (3).....	142

Figure 5-13	Percentage of subjects whose sd of response frequency across stimulus types fell into each shaded range.....	144
Figure 5-14	Total frequency of responses per stimulus type per sentence.....	145
Figure 5-15	Median, interquartile range and minimum/maximum of response frequency per stimulus type.....	147
Figure 5-16	Median, interquartile range and minimum/maximum of difficulty ratings.....	148
Figure 5-17	Difficulty ratings split by musical training category.....	150
Figure 5-18	Criteria used by subjects to judge natural-sounding rhythm.....	150
Figure 5-19	Three most common answers when subjects were asked to explain their technique for the task.....	151
Figure 6-1	Relative weighting of duration and f0 excursion for four PVI types.....	173
Figure 6-2	Mean PVI scores (10 subjects per language).....	177
Figure 7-1	Duration-based rhythm research compared to how future rhythm research should proceed.....	201

Table of tables

Table 2-1	Summary of Swiss German and (Swiss) French prosodic properties.....	50
Table 3-1	Summary of subjects (experiment 1).....	58
Table 3-2	f0 contours and durations of stimuli.....	59
Table 3-3	Three durations of [si] monosyllables.....	60
Table 3-4	Summary of stimulus pairs (trials).....	61
Table 3-5	Mean and standard deviation (across <i>k</i> subjects) of correct ‘filler’ responses...	63
Table 3-6	‘1 st sound is longer than 2 nd sound’ responses, and significance tests.....	64
Table 3-7	‘Dynamic is longer than level’ (‘D>L’) responses, and significance tests.....	65
Table 3-8	Output of regression model (‘D>L’ responses).....	66
Table 3-9	ANOVA output: <i>sound</i> × <i>direction</i> × <i>language</i>	69
Table 3-10	ANOVA output: <i>sound</i> × <i>duration</i> × <i>language</i>	72
Table 3-11	ANOVA output: <i>sound</i> × <i>order</i> × <i>language</i>	74
Table 3-12	‘High>low’ responses, and significance tests.....	75
Table 3-13	Output of regression model (‘high>low’ responses).....	76
Table 4-1	Examples of labels for variously-sized prosodic groups.....	83
Table 4-2	Summary of subjects (experiment 2).....	86
Table 4-3	Summary of bilinguals’ language backgrounds.....	88
Table 4-4	Syllable sequences chosen for stimuli.....	90
Table 4-5	Base duration (d1) of each syllable.....	94
Table 4-6	Duration of each syllable after lengthening (d2).....	95
Table 4-7	Number of trials.....	98
Table 4-8	Variables in experiment design.....	99
Table 4-9	Frequency of 3+2 responses out of total number of control trials per language group, and significance tests.....	101
Table 4-10	Output of regression model.....	104
Table 4-11	Post-hoc pairwise comparisons for <i>pattern</i>	105
Table 4-12	Planned contrasts for the <i>cue(s)</i> × <i>language</i> interaction.....	107
Table 4-13	Post-hoc pairwise comparisons for <i>cue(s)</i>	108
Table 4-14	Post-hoc pairwise comparisons for <i>cue(s)</i> × <i>language</i> (bilinguals).....	109

Table 4-15	ANOVA output: <i>cue(s) × group</i> (comparing monolinguals and bilinguals).....	110
Table 5-1	Prosodic structure of French stimuli.....	123
Table 5-2	Prosodic structure of Swiss German stimuli.....	124
Table 5-3	Duration and/or f_0 manipulation made in nine conditions.....	125
Table 5-4	Mean percentage of correct responses in AXB task.....	127
Table 5-5	Nine stimulus conditions.....	131
Table 5-6	Summary of subjects (experiment 3).....	136
Table 5-7	ANOVA output: <i>f₀ excursion × duration × language</i>	139
Table 5-8	ANOVA output: <i>f₀ excursion × duration × region</i>	143
Table 5-9	Output of regression model (random factors only).....	146
Table 6-1	Summary of subjects (experiment 4).....	165
Table 6-2	Examples of X_{dur} values depending on original syllable duration.....	169
Table 6-3	Examples of X_{f_0} values depending on original syllable excursion.....	169
Table 6-4	Output from each regression model.....	170
Table 6-5	Explanation of standardised b-coefficients.....	171
Table 6-6	Weighting values used in weighted PVIs.....	172
Table 6-7	How to calculate each PVI (hypothetical data)	175
Table 6-8	Comparison of PVIs for hypothetical languages.....	176
Table 6-9	ANOVA output: <i>interval × PVI-type × language</i>	178
Table 6-10	Planned comparison within ANOVA (<i>interval × PVI-type × language</i>).....	179
Table 6-11	ANOVA output: <i>interval × PVI-type × language</i> (Swiss French vs. French).....	181
Table 6-12	Correlation output.....	182
Table 6-13	Speech-rate-related properties measured from recordings.....	183
Table 7-1	Summary of research questions and answers.....	191
Table 7-2	Models that consider how native language might constrain perception.....	202

List of abbreviations

BiSFr	Bilingual (listening to (Swiss) French)
BiSG	Bilingual (listening to Swiss German)
f_0	fundamental frequency
Fr	French
Frs	French subjects
IP	Intonation Phrase
PVI	Pairwise Variability Index
RG	rhythmic group
SFr	Swiss French
(S)Fr	Swiss French and French
SFrs	Swiss French subjects
(S)Frs	Swiss French and French subjects
SG	Swiss German
SGs	Swiss Germans, Swiss German subjects
SstG	Swiss standard German

Acknowledgments

Although written by one person, this PhD thesis had many more people involved behind the scenes, without whose help it would not have been possible. I hope that the following words of thanks will have left no-one out.

I am convinced that the kind of research which confines the researcher to sitting alone in a library or at home would have driven me to despair. Instead I was fortunate enough to have a ‘workplace’ where isolation was out of the question. I always looked forward to entering room 219 of the Raised Faculty Building, where my office-mates were present, in body or spirit. Thanks to Marco for not minding when papers and other (non-)academic clutter spilled over onto his desk, and to Spyros for amusing computer sound-effects and helpful explanation about statistics. Most notably, thanks to Hae-Sung, not only for doing research similar enough for us to have rhythm-related rants and to give me feedback on this thesis, but also for being a friend beyond our rather long office hours. Just along the corridor was the Phonetics Lab, where I found yet more support in the form of a crowd around the kettle; chatting over tea could improve the most frustrating of days. Thanks to everyone present, particularly Meg for providing such amazing cakes and for giving feedback on this thesis, Antje for her patience in answering my statistical questions, and Geoff for never being too busy to provide technical support. I hope to have completed a thesis with which my supervisor, Francis, is pleased. I am thankful for his idea to pursue the topic of non-durational aspects in rhythm, which I thoroughly enjoyed researching. Under his expert guidance, I always felt that the next step at every stage along the path through research was clear.

Most fieldwork took place in Switzerland. This would have been a bigger challenge without the kindness of several people. Thanks to Stephan Schmid for warmly welcoming me into Zürich University Phonetics Lab and giving me work space, especially the booth. Likewise thanks to Daniel Elmiger in Neuchâtel for his generous hospitality at the university, by reserving a testing room, and to Sandra Schwab for help with finding Swiss French participants. For Swiss German, thanks to Niels Anner and Lea Hagmann for help with recruiting participants and stimulus preparation. I am also grateful to the Hagmann family for making me feel at home during my stays in Zürich. Many bilingual participants were recruited thanks to Katia Rüeeggsegger. The participants themselves also deserve my thanks.

I would never have got to the position of being able to do a PhD without my parents, who always encouraged me to achieve to my full potential, and provided generous funding throughout my education. It is difficult to pay them back with simply words. Then last, but by absolutely no means least, I will never be able to put into words how thankful I am to my husband, Tom. Without him, I would not have remained as motivated as I did throughout various stages of the work. At least this thesis bears his name too, and I dedicate it to him.

Introduction

1.1 What is rhythm?

Rhythm is a fundamental part of life. In humans, our ability to produce and perceive rhythmic behaviour, e.g. in music and language, might have evolved from primates (Fitch et al. 2005), or might be uniquely human (Pinker and Jackendoff 2005). Non-experts would probably recognise the term ‘rhythm’, yet scholarly writing on the phenomenon since ancient times has not provided one simple definition, for two main reasons. First, rhythm manifests itself in various forms (Adams 1979). In the arts, the term is used in several fields including music, e.g. simple or compound rhythm, and poetry, e.g. iambic rhythm. In science, it is applied to: anatomical functions, e.g. respiratory rhythm, locomotor activity rhythm, heart (sinus, cardiac) rhythm; regularly-recurring phenomena, e.g. (daily) circadian rhythm; and processes in the natural world, e.g. tidal rhythm, rock formation rhythm. Second, the concept of rhythm to which researchers in each field refer is complex with several variables (see e.g. Fraisse 1982).

Therefore, the above question does not have a simple answer. It is, however, essential to define what rhythm is in the context of this thesis. First, in the following three definitions from dictionaries, one general followed by two linguistic, notice that ‘regularity’ is important (emphasis RC):

Rhythm. [general:] Movement marked by the **regulated succession** of strong and weak elements, or of opposite or different conditions.’ (*Oxford English Dictionary* 1989)

Rhythm. ‘An application of the general sense of this term in phonology, to refer to the **perceived regularity** of prominent units in speech. These **regularities** may be stated in terms of patterns of stressed v. unstressed syllables, syllable length (long v. short) or pitch (high v. low), or some combination of these variables.’ (Crystal 1985: 266-67)

Rhythm. ‘The perceptual pattern produced in speech or poetry by the occurrence at **regular intervals** of prominent elements; these elements may be stresses (as in English), syllables (as in Spanish), heavy syllables (as in Ancient Greek) or moras (as in Japanese).’ (Trask 1996: 311)

In psychology, Brown (1908: 336) summarised the views of his contemporaries researching rhythm by stating: ‘Some hold that this impression [of regularity] arises from the regular recurrence, in **time**, of certain features of the rhythmic series; others claim that the regularity resides in the **structure** of the elements composing the series’ (emphasis RC). Crystal (1969) and Adams (1979), writing specifically about linguistic rhythm, labelled these two views as ‘timer’ and ‘stresser’ respectively: research focused on either the timing of, or the

acoustic/phonological structure of, prominent and non-prominent units in the speech signal/written text. This thesis presents experimental phonetic research on rhythm in the acoustic speech signal from a ‘stresser’ perspective, unlike many recent phonetic experiments on rhythm that have a ‘timer’ perspective.

As a working definition, speech rhythm is:

- (i) the **perceived** regularity in an utterance,
- (ii) induced by the acoustic **multi-dimensionality** of the speech signal (which results from a realisation of phonological structure),
- (iii) and influenced by the listener’s **native language**.

The originality of the research in this thesis relates to the bolded words, to which reference will later be made when this definition is contextualised within current experimental phonetic research on speech rhythm; how this research area developed into its current state is traced in the following sections.

1.2 Review of speech-rhythm research

Given that rhythm is a multidisciplinary research topic, only a small proportion of publications on rhythm are in linguistics. A search through journals published in the years 2007 to 2010 for articles concerning rhythm gives a range from biological and physical sciences including medicine, neuroscience, psychology, engineering and computer science, to arts and humanities including music, literature, poetry, philosophy and linguistics. The following review comprises mostly linguistic research on spoken language, and traces the development of experimental phonetic research on rhythm. For clarity, henceforth the term ‘rhythm’ means ‘speech rhythm’, unless otherwise specified. However, the review includes disciplines other than phonetics which were precursors to phonetic experiments on rhythm, or will be important when, after the review, we consider that current experimental phonetic research on rhythm is deficient in three areas of activity:

- (1) empirically investigating perception (from a non-typological perspective),
- (2) measuring acoustic cues other than duration,
- (3) testing whether native language/dialect influences speakers’ perception of rhythm.

Notice that these three points resemble the working definition above. The review consists of five chronological sections, each covering a (longer or shorter) era in which certain lines of research dominated. Relatively more words are devoted to the current era (twenty-first century), as this research has had less opportunity to appear in historical reviews. Detailed histories of rhythm research also appear in Adams (1979), Pamies Bertrán (1999) and Kohler (2009a).

1.2.1 Pre-1900

Ancient Greek philosophers identified the phenomenon of *ῥυθμός* ('rhythmus') meaning 'a measured movement' (Partridge 1961), whence the English term 'rhythm'. According to Lucas (1968), Aristotle saw rhythm as 'a pattern of recurrence imposed on speech or on other sounds' (Adams 1979: 10). Aristotle's pupil Aristoxenus later added the idea that the timing of the recurrence was important (Adams 1979: 10).

Many centuries later, Steele (1775) investigated rhythm and melody in English. He analysed speech into 'cadences', using notation similar to a modern musical stave with notes and bars representing syllables and cadences. Each cadence had one 'heavy' and at least one 'light' syllable, forming an alternating pattern of heavy–light etc., or *arsis–thesis* (from Greek which he deemed more pleasant than 'vulgar' English) (Steele 1775: 18, 22). Two points about his work later became important. First, the sum of syllable weights was equal in each cadence, so cadences were similar in duration. Second, Steele (1775: 120) admitted that rhythm was not identical in English and Ancient Greek, and briefly noted that the English word *impossible* becomes the French word *impossible* if the heavy and light beats change location.

1.2.2 1900-1939

This era saw an increase in interest for investigating rhythm with experiments, particularly amongst psychologists (see also Spitznagel's (2000) summary) and some phoneticians.

1.2.2.1 Psychology

Several psychologists investigated how listeners perceived rhythm in series of non-speech sounds (e.g. Bolton 1894, Wallin 1911, Woodrow 1909). Others were interested in language, and attempted experiments whereby rhythm perception was investigated through production of motor movements. Brücke (1871) used a marker on a smoked drum to record a subject's finger movements tapping to stressed syllables whilst Brücke recited German verses; he found that inter-tap intervals (representing feet) were durationally equal (Scripture 1902: 537). Wallin (1901) reviewed several experiments similar to Brücke's (1871), and concluded that rhythm investigation would be improved if indirect, irreproducible measurements like tapping were replaced by direct measurements of sounds recorded on, and reproduced by, a phonograph. Likewise Brown (1908: 13) held previous methods (like tapping) not 'scientific' enough for the 'scientifically minded psychologist'.

From phonograph recordings, Wallin (1901) measured the duration of feet and lines of verse in several English speakers' poetry reading; he perceived utterances with not exactly durationally equal feet and lines as rhythmical. From kymograph recordings of laryngeal vibrations and supralaryngeal movements, Brown (1908) determined syllable and foot durations in verse (meaningful and nonsense [pa]-syllable series) read by English speakers. Feet were not

durationally equal in meaningful verse, though tended towards equality in nonsense verse, from which Brown (1908: 71) concluded that spoken verse does not have regular timing as others had claimed from auditory impressions. Nevertheless, Brown (1911: 343) later argued that although verse rhythm is irregular, 'it is of a very distinctly temporal type, giving, all the disturbing factors being considered, very great regularity in the matter of recurrence'. Wallin's and Brown's experiments moved rhythm investigation from perception to production (though note Wallin's (1911) later perceptual work on non-speech).

1.2.2.2 Phonetics

As general phonetics manuals became more prevalent, they often included notes on rhythm (e.g. Jones 1956: first published 1918, Rousselot 1924, Scripture 1902). Jones (1956: 239, 242-44) admitted his lack of time to investigate rhythm in detail, but briefly observed that English has 'a general tendency to make the "stress-points" of stressed syllables follow each other at equal intervals of time', and that French learners speak English with durationally similar syllables. Rousselot (1924: 1094) briefly reported a recording of a French speaker's verse reading, and concluded that syllables were almost durationally equal when produced in isolation, but in longer utterances, timing regularity became important at the foot level. Scripture (1902) devoted two chapters to rhythm (general and speech), perhaps due to his interest in rhythm from earlier work in psychology (e.g. Scripture 1897, 1899). From his phonograph-based experiments on French and English verse, he concluded, contrary to previous authors, that feet did not exist in speech, and suggested that a 2:1 timing relationship between heavy and light syllables (*arsis-thesis*) occurred irregularly and by chance (Scripture 1902: 552-54).

Monographs on rhythm from a phonetic perspective also started to appear. Sievers' (1912) book on rhythm and melody in spoken German verse was based on his auditory impressions, but he concluded that rhythm research should move from subjective commentary to objective experiments, as Classe (1939) did for his monograph. Classe (1939) chose not to investigate rhythm perception, because it had already been studied more than production, which he wanted to approach from a phonetic perspective. Using kymograph records of speakers reading English text passages, he measured durational properties, concluding that an 'English sentence is normally composed of a number of more or less isochronous groups, which include a varying number of syllables' (Classe 1939: 132).

1.2.2.3 Summary (1900-1939)

In this era, rhythm research moved from auditory observations to objective investigations. Speech recording and measuring instruments were relatively primitive, and speakers mainly produced read verse with specific metrical structure. The languages tested were often English, German or French, though perhaps this work was and is most accessible to later

rhythm researchers, many of whom have published in these languages. For Spanish, Pamies Bertrán (1999) reported some duration measurement work by Navarro Tomás (1916, 1917 etc.).

1.2.3 1940s-1970s

This era saw the concept of syllable or foot ‘isochrony’ become a hot topic. Traces of this idea had appeared in Steele’s (1775) auditory observation, and some experiments had begun to test this (e.g. Brown 1908, Brücke 1871, Classe 1939, Wallin 1911).

1.2.3.1 Rhythm typology

What is now called ‘rhythm typology’ emerged from the concept of isochrony in speech. Lloyd James (1940) referred to two types of speech rhythm, ‘machine-gun’ and ‘Morse-code’, which lost their wartime connotations when Pike (1945) proposed the terms ‘syllable-timed’ and ‘stress-timed’ respectively. According to Pike’s (1945: 34) auditory impression, though he quoted Classe’s (1939) experimental findings, in ‘syllable-timed’ and ‘stress-timed’ speech the syllables and stresses respectively ‘*tend to come at more-or-less evenly recurrent intervals*’ (emphasis RC). Note that he did not suggest strict isochrony. He also implied that two rhythm types can exist in one language, since English is usually ‘stress-timed’, though in some circumstances (e.g. chanting) it can be ‘syllable-timed’ (Pike 1945: 35). He used these terms in the limited context of American English spontaneous speech, as the purpose of his work was to assist non-native speakers having problems speaking English with native-like prosody (Pike 1945: 1). Although some notion of speech isochrony had already existed for some time, others then attributed it to Pike (1945), not always accurately interpreting his words. For example, Whitehall and Hill (1958: 396) credited Pike with discovering that ‘isochronism’¹, the time between two primary stresses being equal, is a feature of spoken English.

In an assertion which became highly influential in rhythm research, Abercrombie (1967: 97, fn 7) cited the impressions of Lloyd James, Pike and ‘many writers’, and transformed their ideas into a rhythm typology. On two pages about rhythm in his general phonetics textbook, Abercrombie (1967: 97) asserted that all the world’s languages could be dichotomously classified as ‘stress-timed’ or ‘syllable-timed’. He described the inter-stress/-syllable intervals in either rhythm type as ‘isochronous’ (without ‘more-or-less’ like Classe (1939) and Pike (1945)). As examples, Abercrombie (1967) gave French, Telugu and Yoruba (syllable-timed) and English, Russian and Arabic (stress-timed). During a 1971 lecture series (reported in Adams 1979: 52), Abercrombie stated that ‘[u]sually a language has one or the other type of rhythm but not both since the two types are incompatible. [...] English has a stress-timed rhythm which manifests

¹ According to the *Oxford English Dictionary* (1989), *isochrony* and *isochronism* are synonyms, but *isochronism* dates from 1770 (around the time of Steele) and *isochrony* dates from 1953.

itself in all modes of spoken expression.’ This opposed Pike’s (1945) idea of one language having different rhythms depending on speech style.

Abercrombie (1967) did not mention previous writing on Japanese rhythm and contemporaneous experiments on English isochrony. According to Warner and Arai (2001), writers before Abercrombie (e.g. Bloch 1950, Jinbo 1980: originally published 1927, Trubetzkoy 1958: originally published 1939) had claimed that Japanese rhythm has (roughly) isochronous repetition of ‘moras’ (not syllables or stresses). Yet ‘mora-timed’ languages did not feature in Abercrombie’s rhythm typology. (Several experiments in the 1980s-90s led some researchers to propose a mora-timing rhythm type (see Warner and Arai 2001).) In the 1960s, several studies (e.g. Bolinger 1965, Duckworth 1967: cited by Adams 1979, O’Connor 1965, Shen and Peterson 1962) measured inter-stress durations in recorded English utterances and found little or no evidence for physical isochrony of stresses like Abercrombie (1967) claimed. Experiments which Abercrombie (1967: 35, fn 4) did mention were those by Ladefoged and colleagues. They investigated the respiratory muscles’ activity using electromyography, and found no evidence for Stetson’s earlier claims that each syllable is accompanied by a ‘chest pulse’ of rib muscle activity and that abdominal muscles reinforce this activity in each stressed syllable (Ladefoged et al. 1958, and other studies reported in Ladefoged 1967). However, Abercrombie (1967: 36, 96), citing Stetson’s earlier work, claimed that speech rhythm is determined by muscular activity controlling ‘chest-pulses’ and ‘stress-pulses’ in the air-stream mechanism. According to Kelly’s (1993) obituary for Abercrombie, in his defence, his rhythm statement was apparently misinterpreted, and generally he emphasised that laboratory measurements were meaningless without relation to language functioning.

1.2.3.2 Isochrony investigation

Of the 1960s investigations of English isochrony mentioned above, Bolinger’s (1965) is noteworthy (for later discussion: §1.2.6.2) because it considered pitch as well as timing. Bolinger (1965) suggested that spoken discourse lacks inter-stress isochrony because syllable length is intimately related to pitch accent, whose placement can change in spontaneous utterances depending on several linguistic factors. In the 1970s, some phoneticians (and some psychologists), accepting the evidence that ‘objective’ isochrony was absent, investigated whether ‘subjective’ isochrony was present in English (Allen 1972, Donovan and Darwin 1979, Fowler 1979, Lehiste 1973, 1977, Morton et al. 1976). (Chapter 5 reviews some of these studies’ methodologies; their relevance to subjective/perceptual isochrony is discussed here.) Classe (1939: 133) had already suggested that ‘[i]n speech, long groups, [...] will tend to be made subjectively isochronous by the reader or listener because of his speech habits.’

Lehiste (1973) recorded two speakers reading four-foot sentences, and oscillographic measurements showed that the feet were not physically isochronous. These sentences were then

used in a perception experiment, in which listeners were unable to consistently differentiate longer from shorter feet, from which Lehiste (1973) concluded that listeners must hear these rhythmic units as durationally equal in some sense. In the same experiment with non-speech sounds, listeners found it much easier to consistently indicate the longest and shortest out of four noise-filled intervals, as Classe (1939: 88) had anecdotally noted. From these and other contemporaneous findings, Lehiste (1977) concluded that speech is perceptually isochronous, and that the durational differences between inter-stress intervals may be below the ‘just noticeable difference’ threshold for speech sounds. This is only one possible explanation, as this thesis will explore. Donovan and Darwin (1979) concluded similarly to Lehiste (1977) that isochrony is perceptual and confined to language; they found that listeners perceived speech, but not non-speech, as more isochronous than the acoustic signal when listeners adjusted noise bursts to match the rhythm of sentences or tone sequences, or tapped to stressed syllables in four-foot sentences.

Allen (1972: 190) argued, from the results of several experiments he ran, that listeners decode the stress-timed rhythmic structure of English speech by perceiving ‘stress beats’. The location of stress beats was indicated by the time at which subjects tapped or placed a non-speech click on stressed syllables; many indicated beat location around the vowel onset, though with relatively high inter-subject variation (Allen 1972: 82, 92). Morton et al. (1976) suggested that this variation resulted from tasks requiring absolute (not relative) beat location; their listeners determined the timing of syllables relative to others, and the results demonstrated that a speech sound’s ‘psychological moment of occurrence’ or ‘P-centre’² (‘Perceptual-Centre’, i.e. beat) is not its onset (Morton et al. 1976: 405-08). Fowler (1979: 377) instructed a speaker to say nonsense monosyllables ‘at a slow, rhythmic rate, stressing every syllable’, and found that the deviations from acoustic isochrony that he produced were precisely like those which Morton et al.’s (1976) listeners required for perceived isochrony. When Fowler (1979) manipulated these productions to be physically isochronous, and presented them in a perceptual task, listeners judged the non-manipulated original productions as more ‘rhythmic’ than the physically isochronous stimuli. Fowler’s (1979) subsequent production studies found that P-centre location might not correspond to superficially invariant (i.e. measurable) acoustic properties like intensity or fundamental frequency (henceforth f_0) peaks. All these studies on perceptual isochrony demonstrated that measuring inter-stress intervals from acoustically defined points like syllable onset/offset did not reflect perceived timing.

² Some authors use the American spelling ‘P-center’; I use the British spelling. According to Morton et al. (1976: 405), these spellings are synonymous.

1.2.3.3 Non-isochrony-based research

Some lines of rhythm research in the 1970s did not primarily concern isochrony. These included articulatory investigation (e.g. Stone 1979), psychological research on linguistic rhythm (e.g. Martin 1972), phonetic studies on rhythm acquisition, and a rhythm-based phonological theory. The latter two, which link to later work, are discussed briefly.

Allen and Hawkins (1979, 1980) observed from conversational recordings of English-speaking two- to three-year olds that the older the child, the more reduced syllables they produced, and the more adult-like their use of phonetic stress-signalling properties. This suggested that children's rhythm was initially 'syllable-timed' and later became more 'stress-timed', though Allen and Hawkins (1980: 250) admitted that this dichotomy was simplistic and that more research was needed on rhythm generally. In perceptual experiments, Jusczyk and Thompson (1978) and Spring and Dale (1977) demonstrated that one- to four-month-old infants could discriminate changes in stress pattern, so were already sensitive to rhythmic cues (Echols et al. 1997: 204). Research on first language rhythm acquisition increased greatly in the 1980s-1990s. Adams (1979), interested in second language (L2) rhythm, recorded native Australian English speakers and L2 English speakers whose native languages were reportedly 'syllable-timed'. Electromyographic and acoustic analysis revealed little difference between the groups' productions of stress correlates. Adams (1979: 155) concluded that the difference between native and non-native rhythm lay in the timing of rhythmic units.

Lieberman (1975) proposed and developed (in e.g. Liberman and Prince 1977) the theory of Metrical Phonology, which specifically concerned rhythm (Ladd 1996: 205). In this theory, prominence and timing were partly independent: Liberman (1975) differentiated between metrical trees (expressing strong-weak abstract hierarchical organisation) and metrical grids (modelling temporal elements of trees) (Couper-Kuhlen 1993: 86). This theory aroused interest amongst phonologists through to the 1980s over rhythm's role in phonological systems (Echols et al. 1997, who cited as examples: Halle and Vergnaud 1987, Hayes 1981, Prince 1983, Selkirk 1984).

1.2.3.4 Summary (1940s-1970s)

In this era, a rhythm typology emerged, and the number of phonetic experiments on rhythm increased, as speech-recording equipment developed in accuracy and availability. Isochrony was the greatest concern, almost exclusively in terms of English 'stress-timing'; perceptual and physical isochrony were differentiated.

1.2.4 1980s-1990s

This era saw even more phonetic experiments on rhythm, mainly related to isochrony and rhythm typology, both of which were highly controversial. Languages other than English were investigated.

1.2.4.1 Stress-timing/syllable-timing rejected

Several publications presented duration measurements to demonstrate the lack of physical isochrony for: stresses in so-called ‘stress-timed’ English (e.g. Dauer 1983, Faure et al. 1980, Jassem et al. 1984, Roach 1982); syllables in so-called ‘syllable-timed’ French (e.g. Dauer 1983, Roach 1982, Wenk and Wioland 1982) or Spanish (e.g. Pointon 1980); moras in so-called ‘mora-timed’ Japanese (Hoequist 1983a, 1983b). Fletcher (1991: 195) was concerned about the assumption that syllables were the rhythmic unit in ‘non-stress-timed’ languages, so she measured durational properties of French prosodic groups comparable to English feet, and found non-isochronous syllables, and longer-domain timing patterns similar to other languages. Similarly, Pamies Bertrán (1999) measured syllable *and* inter-stress-interval durations in several languages (some so-called ‘stress-timed’, some so-called ‘syllable-timed’), none of which showed stress or syllable isochrony. He argued that the stress-timed/syllable-timed dichotomy had emerged from the application of ancient simplistic definitions of rhythm to the complexity of speech, and suggested that linguists should be open to the possibilities that polyrhythmic structures may exist in some languages, and no rhythm in others.

Miller (1984) investigated whether ‘stress-timing’ and ‘syllable-timing’ were perceptually justified. After hearing short extracts of read and conversational speech, listeners (native English and French speakers, both phoneticians and untrained subjects) inconsistently judged various languages as stress-timed or syllable-timed. Miller (1984) interpreted these results as evidence that both rhythmic types exist in single languages, though admitted the problems with this experiment: untrained listeners need (unbiased) explanation of rhythm types; phoneticians are biased by their linguistic training.

1.2.4.2 Alternative rhythm typologies: categorical or continuous

Wenk and Wioland (1982) suggested ‘leader-timed’ and ‘trailer-timed’ to describe English and French rhythm respectively, but doubted that all languages could be classified with this dichotomy. Similarly, Vaissière (1983: 64, 1991: 118) suggested ‘stress language’ and ‘boundary language’ for English and French respectively. (Chapter 2 gives more detail on these suggestions.) Hoequist (1983b: 229) argued that a stress-timed/syllable-timed typology is only tenable if we admit non-strict isochrony, and suggested that languages could be grouped as ‘duration-controlling’ (e.g. Japanese), ‘duration-compensating’ (e.g. English) or neither (e.g. Spanish), based on Fowler’s (1977) terminology. Instead of a dichotomous typology, Dauer (1983, 1987)

proposed a rhythm-type continuum along which languages are more or less ‘stress-based’³ depending on their linguistic structure. Dauer (1983), after observing various characteristics of languages in which she found no physical isochrony, suggested that perceived rhythmic differences result from language-specific phonological, phonetic, syntactic and lexical properties, e.g. syllable structure, magnitude of vowel reduction, and existence and phonetic nature of lexical stress. Two contemporaneous studies on other languages (Strangert 1985: Swedish, Engstrand 1987: Sami) concluded similarly (see Engstrand and Krull 2002).

Dauer (1987) proposed a ‘rhythm score’ system, whereby each language could be scored qualitatively as to which one of two/three characteristics it displayed for eight rhythm-related phonological properties. Two separate research groups sought a method for quantifying from the acoustic signal these qualitative properties that characterised languages’ rhythms. One was initiated by Low’s (1994) MPhil research, for which Nolan proposed the ‘Pairwise Variability Index’ (PVI), which calculated the difference in a given acoustic property between the members of each successive pair of vowels, and from this the mean pairwise difference across an utterance. Low (1994), Low and Grabe (1995), Low (1998), and Low et al. (2000) investigated Singapore English (reportedly more syllable-timed than British English); the smaller extent of vowel reduction in Singapore English than British English was reflected in the values obtained from PVIs of duration, amplitude and spectral dispersion measurements. The other research group interested in rhythm quantification was Ramus and colleagues, whose work is discussed below with reference to infant studies (§1.2.4.5).

Like Dauer (1983), Auer (1993) concluded, after surveying thirty-four languages’ properties from their phonological descriptions, that prosodic typology is a continuum. Auer (1993: 12) claimed that ‘[p]ossibly the best-known attempt to devise a prosodic typology which includes the (phonologically revised/reinterpreted) distinction between stress-timing and syllable-timing comes from Donegan & Stampe (1983).’ This may be true for phonology, but Donegan and Stampe (1983) were not widely cited within phonetic research. Auer (1993: 95) argued that a timing-based typology was less appropriate than one based on how many phonological rules refer to each prosodic unit (from mora to Intonation Phrase, henceforth IP) in different languages. Auer (1993: 90-95) proposed that in many languages most rules refer either to the syllable or to the phonological word, hence ‘syllable languages’ and ‘word languages’. These lie at either end of the continuum, and non-prototypical languages, with several rules referring to other units, lie in between. However, Auer’s later work on rhythm in turn-taking concerned timing. From measurements of stress-group durations in English and German conversational speech, Auer and Couper-Kuhlen (1994), Couper-Kuhlen (1993), and Auer et al. (1999) concluded that rhythm is

³ O’Connor (1973) and Allen (1975) had used ‘stress-based’ and ‘syllable-based’.

an interactive phenomenon related to meaning in context, since speakers coordinated rhythmic beats (i.e. maintained rhythmic structure) at turn onsets. Auer and Couper-Kuhlen (1994: 103) and Couper-Kuhlen (1993: 297) suspected from preliminary evidence that speakers of rhythmically different languages might behave similarly at turn-taking, since all have interactional concerns. (Further evidence for this was later provided by Szczepek Reed (2010) from conversations in English between a native speaker and speakers whose native language was reportedly syllable-timed.) Unlike in most phonetic research on rhythm, Auer and colleagues, like Bolinger (1965), stressed the need to investigate spontaneous speech.

1.2.4.3 Perceptual isochrony and P-centres

Scott et al. (1985) were concerned that Donovan and Darwin's (1979) tapping experiment (§1.2.3.2) only investigated English, so they constructed equivalent sentence-sets in English and French, and presented each set to both English and French subjects, who had to tap on the four stressed syllables. Both subject groups tapped more isochronously to French than English sentences, though the French were overall slower tappers than the English. Scott et al. (1985) concluded that English speakers' isochronous tapping was not evidence that English has an underlying stress-based isochrony, unless French does too. From a subsequent experiment, Scott et al. (1985: 161) argued (with an interpretation different from Donovan and Darwin's) that subjects may simply be biased toward isochronous taps when the task becomes increasingly spectrally complex, because subjects tapped more isochronously (than the timing of the stimuli) to segmentally degraded sentences and intelligible sentences, but not to noise-burst sequences.

Benguerel and D'Arcy (1986) presented English, French and Japanese listeners with stimuli comprising six clicks or syllables which were manipulated to accelerate or decelerate to various extents; all language groups perceived stimuli within a certain range of (mainly decelerating) manipulations as isochronous. Benguerel and D'Arcy (1986: 244) suggested that pre-production speech is intended to be isochronous, but becomes temporally disrupted due to articulatory and linguistic constraints. This links to Vatikiotis-Bateson and Kelso's (1993) articulation model (based on rhythm experiments with English, French and Japanese speakers), in which underlying universal parameter values are set language-specifically depending on constraints like syllable structure in production, and maintaining temporal distinctions in perception. (See also Classe's (1939: 100) idea that perfect isochrony can only be realised when intervals have similar phonetic and syntactic structures.)

In the 1980s, P-centres in speech (Morton et al. 1976) became a popular line of research in phonetics and psychology; most studies concerned English. Tuller and Fowler (1980) conducted a production experiment in which English speakers produced series of monosyllables as if in time with a metronome, whilst electromyographic potentials were recorded from their lip muscles. P-centres (i.e. the time-points at which speakers produced syllables according to what

they perceived to be isochronous sequences) were found to correlate with isochronous muscular activity, but there was no obvious unique articulatory P-centre correlate (e.g. the vowel gesture's onset). Likewise, Patel et al. (1999) could not determine a unique P-centre correlate by taking various acoustic and articulatory measurements from a similar production task. In a similar experiment, Hoequist (1983c) tested English, Spanish and Japanese speakers; in all three groups, P-centres lay somewhere between syllable onset and periodicity onset. Hoequist (1983c) concluded that P-centres are probably universal and not characteristic of a particular rhythm type, but might relate to language-specific units (i.e. moras, syllables, stressed syllables in mora-/syllable-/stress-timed languages respectively).

Others conducted perceptual experiments, in which English listeners adjusted via a knob the timing of alternate monosyllables until they perceived each pair to occur at an equal interval; various acoustic properties of syllables were manipulated to investigate how they influenced P-centre location. These experiments found that segmental durations within the syllables influenced P-centre location. However, some of these experiments found that amplitude properties, like within-syllable energy distribution and the rise time of the syllable-onset envelope, were just as influential as, or more so than, segmental durations in determining P-centre location (Harsin 1997, Howell 1984, cited in Howell 1988, Pompino-Marschall 1989, Scott 1994, 1998), whereas others found that amplitude properties influenced P-centre location much less than duration did (Marcus 1981, Cooper et al. 1986). Fox and Lehiste's (1987) methodology differed slightly from the perceived interval task; they found that increased vowel duration, but not vowel quality, influenced P-centre location. Marcus (1981) and Fox and Lehiste (1987) argued that P-centre location is determined by the acoustic signal across the whole syllable, rather than a specific acoustic event within it. Harsin (1997: 247-51) suggested that 'psychoacoustic processing of spectrotemporal cues' (i.e. spectral properties together with timing) could underlie P-centre location.

Overall, this P-centre research had two important implications for other rhythm research that claimed a lack of isochrony from duration measurements alone. First, physical measurements from acoustically defined points like syllable onset to offset did not reflect perceived isochrony. Second, one unique correlate of P-centres was not found, and the influence of various temporal and spectral properties was tested and debated, implying that speech timing perception is a complex process involving several acoustic cues integrated into each sounds' perceived 'moment of occurrence' (Morton et al. 1976).

1.2.4.4 Phonological models

Several phonologists developed Metrical Phonology, continuing from Liberman (1975), e.g. Selkirk (1984) who demonstrated the 'Principle of Rhythmic Alternation': continuous speech has a succession of strong-weak-strong-weak etc. (or starting with weak) syllables. (Couper-

Kuhlen (1993: 82, fn) noted: '[a]s a notion, rhythmic alternation dates back to Sweet (1875-6) and Jespersen [(1933)] (1970: 254).' In this framework, Arvaniti (1994) proposed that 'stress-timed' languages prefer alternating stressed-unstressed syllables, whereas 'syllable-timed' languages tolerate longer stretches of unstressed syllables, so the rhythm types reflect different hierarchical rhythmic structures. Dasher and Bolinger (1982), referring to *Metrical Phonology*, argued that phonetic properties of the rhythm we hear result from several interacting language-specific phonological factors (cf. Dauer 1983), including pitch accents as Bolinger (1965) had argued, rather than one underlying element specified as 'stress-timed' or 'syllable-timed', so a dichotomous typology was improbable.

Temporal Phonology, another rhythm-based phonological model, was developed by Port, Cummins and colleagues after they investigated Japanese and English. Port et al. (1995: 15) proposed an 'adaptive oscillator model', which modelled rhythm as the hierarchical organisation of temporally coordinated prosodic units. Japanese had a single-level (mora-level) oscillator structure, whereas English had a hierarchical structure with the syllable-level oscillator coupled to the foot-level oscillator (Cummins and Port 1998, Port et al. 1995). This idea of hierarchical rhythmic structure, already discussed by e.g. Martin (1972), continued into the twenty-first century.

1.2.4.5 Infant studies

Research on rhythm in first language acquisition was started in the 1970s (see e.g. Allen and Hawkins 1979, 1980, Crystal 1970, 1973) and was later developed, particularly as techniques for testing young infants' speech perception became well established, e.g. the 'high amplitude sucking', 'heart-rate' and 'visually reinforced infant speech discrimination' paradigms (Ingram 1989: 87-88, Jusczyk 1997: 233-50). Some examples of infant rhythm perception studies from the 1980s-90s include: Bertocini et al. 1995, Levitt and Wang 1991, Mehler et al. 1988, Morrongiello 1984, Nazzi 1997, Nazzi et al. 1998, Sansavini 1997, Trehub and Thorpe 1989. Nazzi and colleagues' research is most relevant to this review of how experimental phonetic research on rhythm developed.

Nazzi et al. (1998) presented newborns from French-speaking families with low-pass filtered stimuli (i.e. prosodic but no segmental information present) from various languages. These infants could discriminate rhythmically different English and Japanese, but not rhythmically similar English and Dutch, and could discriminate Italian and Spanish from English and Dutch. Nazzi et al. (1998) agreed with e.g. Dauer's (1983) claim that language-specific phonological structure results in languages of different rhythm types having different rhythmic-timing acoustic properties, and they argued that these rhythmic properties must be perceptible, since infants in their study could abstract these properties, and on this basis discriminate stimuli. Ramus and Mehler (1999) found that French adults could also discriminate English and Japanese

segmentally degraded stimuli, and concluded that this methodology had further application in testing the validity of rhythm typologies. Nazzi et al. (1998) related their infant findings to Cutler, Mehler and colleagues' studies on word segmentation strategies for lexical access. Mehler et al. (1981) showed that French adults' segmentation strategy was syllable-based. Cutler et al. (1986) did not replicate these results with English adults, but found that their segmentation strategy was stress-based (Cutler and Butterfield 1992, Cutler and Norris 1988, Smith et al. 1989). Cutler et al. (1986, 1992) concluded that a syllable-based strategy was efficient for French, which has syllable-based rhythm, but not English, which has stress-based rhythm, though they admitted that this rhythm dichotomy was simplistic. Nazzi et al. (1998: 763) argued that their newborns demonstrated an early attention to language-specific rhythmic cues, since they discriminated utterances based on the same rhythm types found to underlie adults' speech segmentation.

From the finding that newborns, with no language-specific knowledge of stress or syllabification, could discriminate languages of different rhythm types, Ramus et al. (1999) reasoned that a viable rhythm theory should not rely on language-specific phonological concepts (e.g. stress, syllable). Ramus et al. (1999) measured segmental durations in read sentences from eight languages, then calculated the proportion of speech comprised by vocalic intervals (%V), and the standard deviation of vocalic and consonantal interval durations (ΔV , ΔC). They argued that these metrics accounted for infant discrimination behaviour (found by e.g. Nazzi et al. 1998), and were rhythm-type correlates, since they (particularly %V and ΔC) appropriately quantified the phonetic realisation of phonological properties in languages reported to be stress-timed or syllable-timed.

1.2.4.6 Summary (1980s-1990s)

In this era, many phonetic experiments on produced rhythm found that various languages were not strictly stress-timed or syllable-timed. Alternative theories for rhythmic variation between languages were proposed; some suggested that the phonetic realisation of rhythm results from language-specific phonological structure. Various perceptual experiments demonstrated that spectral and temporal properties influenced perceived timing (which therefore differed from physically measurable timing), and that phonologically naïve infants could discriminate rhythmically different languages using durational cues.

1.2.5 Twenty-first century

This era saw experimental phonetic research on rhythm focus on duration and timing in production (rather than perception).

1.2.5.1 Widespread use of rhythm metrics

Much work was based on the rhythm quantification methods, i.e. rhythm metrics, proposed by Ramus et al. (1999) (%V, ΔV , ΔC) and Low (1994, 1998) (PVI). Ramus et al.'s

(1999) were utterance-global measures, and only concerned duration, whereas the PVI captured rhythmic properties on a local (vowel-pairwise) scale, and was originally applied to various acoustic properties. However, since Grabe and Low's (2002) study, properties other than duration have been forgotten (see chapter 6). For each of eighteen languages, Grabe and Low (2002) calculated durational PVIs from one native speaker's reading of a translationally equivalent text. The resulting plot of each language's consonantal PVI (x-axis) against vocalic PVI (y-axis) generally separated so-called stress-timed and syllable-timed languages, with other languages in between. From this, Grabe and Low (2002) concluded that languages are more or less stress-timed or syllable-timed along a continuum (cf. Dauer 1983). Conversely, Ramus (2002) tentatively interpreted the clustering of languages along the scale of Ramus et al.'s (1999) metrics or PVIs as evidence for categorical rhythm types. The PVI and Ramus et al.'s (1999) metrics were not the first to capture rhythm numerically; others include those presented by Allen (1973), Kozhenvnikov and Chistovic (1965) (amongst several evaluated in Ohala 1975) and Scott et al. (1985)⁴. Nevertheless, Grabe and Low (2002) and Ramus et al. (1999) became seminal papers in what developed into a rhythm-metric research paradigm. Here a selection of the many studies is presented, including the development of these metrics, in statistical formulation and intervals measured, and their application beyond typological questions.

Before Grabe and Low's (2002) publication, Deterding (1994) had suggested a pairwise normalisation component for Low's (1994) original 'raw' PVI formula, to account for confounding effects of speaking rate (Low et al. 2000: 382, fn). Deterding (2001) further adapted the PVI to normalise across the whole utterance (not pairwise). He measured syllable (not segmental) durations, in conversational (not read) speech, and concluded that measuring rhythm like this was appropriate for comparing Singapore English and British English. Without explicit rationale, Gibbon and Gut (2001) removed part of the pairwise normalisation component, and called this formula the 'Rhythm Ratio'. Bertinetto and Bertini (2008) transformed the (raw) PVI into the 'Control/Compensation Index' (CCI), by dividing the duration of each vocalic or consonantal interval by its number of segments to account for phonotactic complexity, which was a core feature distinguishing 'controlling' from 'compensating' languages. Bertinetto and Bertini (2008) proposed that rhythmic differences result from some languages 'controlling' the articulatory/gestural effort per segment, hence similar segmental durations, and some languages 'compensating' through gestural overlap for greater effort in stress with reduced effort in unstressed vowels, hence less even duration (cf. Bertinetto and Fowler 1989, Hoequist 1983b). They saw a resemblance between their work and Port and Cummins' (e.g. 1998), which modelled articulatory gestures by integrating oscillators at different prosodic levels.

⁴ Benguerel (1986) believed Scott et al.'s measure was 'flawed', which they defended (Scott et al. 1986).

Other researchers modified Ramus et al.'s (1999) metrics. Dellwo (2006) introduced VarcoC (ΔC divided by the mean consonantal interval duration, hence a normalisation), which more successfully than ΔC distinguished English and German from French at all speech rates (White and Mattys 2007b: 238). From Brazilian and European Portuguese data, Duarte et al. (2001) suggested that the median absolute deviation (MAD) of segment durations was a more robust dispersion measure than the standard deviation. From recordings of four languages, Steiner (2004, 2005) computed Δ and % duration statistics separately for vowels, approximants, laterals, nasals, fricatives and stops, and found that the proportion of laterals (%L) and nasals (%N) more successfully separated so-called stress-timed from syllable-timed languages than %V and ΔC .

Recall Ramus et al.'s (1999) argument that phonetic units, perceived even by newborns, should be measured rather than phonological units. Only a few subsequent rhythm-metric studies questioned this. Wagner and Dellwo (2004) measured syllable durations for their PVI-inspired metric called YARD (Yet Another Rhythm Determination), and found evidence for isochrony at different prosodic levels in different languages. In English and German, roughly isochronous groups were bi-/tri-syllabic, whereas in French and Italian they were at least three/four syllables long. Wagner (2007) presented a visualisation method to detect at which prosodic level rhythm appears in a language, by plotting each syllable's duration⁵ (i) (x-axis) against the consecutive syllable's duration ($i+1$) (y-axis), colour-coded depending on the stressed/unstressed/phrase-final nature of $i+1$. Through this method, English and German demonstrated foot-based rhythm (alternating stressed-unstressed syllables), whereas French demonstrated a more phrase-global rhythm, whilst Italian demonstrated both an alternation of stressed and unstressed syllables, and a tendency towards global regularity in unstressed syllables. Similarly, Nolan and Asu (2009), by calculating syllable and foot (durational) PVIs, provided evidence that languages can have co-existing independent rhythms at different prosodic levels. English was less durationally variable at the foot than syllable level, Mexican Spanish less at the syllable level, and Estonian had relatively low variability at the foot and syllable levels. Nolan and Asu (2009) linked their ideas to Cummins and Port's (1998) model with oscillators at different prosodic levels. Barry et al. (2003) proposed a PVI with duration of consonant+vowel (CV) intervals as input, which are similar, but not completely equivalent, to syllables (largely depending on the language). By comparing several rhythm metrics applied to speech in various languages, including regional and speech-style variations, Barry et al. (2003) concluded that the 'PVI-CV' more appropriately distinguished rhythm types than vocalic or consonantal PVIs.

⁵ Syllable durations were 'corrected using P-centres.' (Wagner 2007: 1114)

Other researchers extended Ramus et al.'s (1999) argument for phonetic units. Galves et al. (2002) argued that infants rely on a coarse-grained perception of sonority versus obstruency, rather than a fine-grained vowel versus consonant distinction. Therefore, they proposed two sonority metrics based on spectral frequency information automatically computed from utterances: \bar{S} (the sonority function's sample mean) estimated the proportion of sonorant material (equivalent to %V); δS (the sonority function's variation) estimated the importance of high-obstruency regions (equivalent to ΔC). Galves et al. (2002) concluded that the results obtained from applying their metrics to several languages explained previous findings for infant perception better than Ramus et al.'s (1999) metrics. Volín and Pollák (2009), for the same reason as Galves et al. (2002), calculated metrics (Ramus et al.'s (1999), PVIs) using the duration of high-energy and low-energy intervals automatically extracted from speech using a power-analysis-based computation.

Lee and Todd (2004) and Tilsen and Johnson (2008) also proposed automatically implemented rhythm metrics based on spectral properties, reasoning that prominence-lending cues other than duration were important in rhythm. Lee and Todd (2004) used a 'rhythmogram' algorithm (Todd 1994, Todd and Brown 1996), which computationally segmented speech input, and assigned prominence values (P) to the output, based on f_0 , intensity and duration in the input. Standard deviations of syllabic and sonorant event prominences (ΔP_{syll} , ΔP_{son}) significantly differentiated English from French, and (to a weaker extent) English and Dutch from French and Italian (Lee and Todd 2004). Tilsen and Johnson (2008) proposed the 'rhythm spectrum', based on Fourier analysing the amplitude envelope of bandpass-filtered speech in arbitrarily defined time chunks (around 2.5 seconds for investigating syllable and foot rhythms, longer for phrasal rhythms). From analysing conversational English speech, Tilsen and Johnson (2008: 36) argued that '[t]he presence of periodicity in a variety of frequency ranges shows that [English] speech is rhythmic on stress and syllabic time scales.' Lee and Todd (2004) and Tilsen and Johnson (2008) believed their metrics could complement and improve duration-based rhythm-metric research.

The rhythm-metric studies discussed so far mainly concerned the suitability of such statistics in quantifying rhythm types. The following studies exemplify how metrics were applied in research on (L1 and L2) rhythm acquisition. With the PVI, Grabe, Gut et al. (1999) and Grabe, Post and Watson (1999) investigated French, English and (in the former study) German children's rhythm. French children's PVIs were nearest their mothers', suggesting that French rhythmic structure was the least complex to acquire (Grabe, Post and Watson 1999). In research on the speech style to which language-acquiring children are exposed, Payne et al. (2009) found that English, Spanish and Catalan mothers showed some cross-linguistically distinct rhythm-metric scores for speech directed at their children (aged two, four and six years); in general, though, child-directed speech showed higher %V scores and lower variability in interval

durations (consonantal and vocalic) than in these adults' adult-directed speech. Lleó et al. (2007) found that three-year-old German~Spanish bilinguals' PVI's were similar for their speech in each language, whereas respective monolingual children's PVI's were distinct. Yet PVI results have shown that by adolescence, English~German bilinguals (Whitworth 2002) and Spanish~English bilinguals (Carter 2005) can produce their languages with distinct rhythms. Jeon (2006) and White and Mattys (2007a) evaluated the effectiveness of various rhythm metrics (Ramus et al.'s (1999), PVI's) for revealing potential influences of L1 on L2 rhythm. White and Mattys (2007a), who examined English, Dutch and Spanish learners of the other languages, concluded that %V and VarcoV provided useful insight into L1 influence on L2 rhythm, whereas Jeon (2006), who investigated Korean learners of English, concluded that rhythm metrics have several problems that need attention.

Another application of rhythm metrics, as a clinical tool for dealing with dysarthria, was proposed by Liss et al. (2009), who found that a combination of rhythm metrics successfully distinguished dysarthric from normal speech. The PVI was also applied in musicology (e.g. Patel and Daniele 2003, Patel et al. 2006). In an overview of speech and musical rhythm, Patel (2007) concluded that speech rhythm results from interacting phonological factors, not an organising principle as underlies musical rhythm, but that both disciplines could benefit from rhythm-metric research.

1.2.5.2 Other phonetic research

Most other rhythm-related phonetic experiments in this decade concerned perception. (An exception is Engstrand and Krull (2002, 2003), who demonstrated that speech style and rate affect rhythm production, by measuring CV unit durations in Swedish and Spanish: no metrics were involved.) Jankowski (2001) replicated with French speakers Port et al.'s (1995) finding for English speakers performing a 'speech cycling' task. When repeating a phrase along with a metronome, speakers of French and English (rhythmically diverse languages) tended to place prominence at specific points in the metronome beats' phase cycle, suggesting that some rhythmic processes are not language-specific (Jankowski 2001).

Previous work by Cutler, Mehler and colleagues, which investigated whether (adult/infant) listeners' word segmentation strategy for lexical access depends on native language rhythm type, was continued. Murty et al. (2007) found that Japanese and Telugu-speaking adults both used a moraic strategy for segmenting Japanese stimuli, but Japanese speakers used this strategy more than Telugu speakers for segmenting Telugu stimuli. Kim et al. (2008) found that Korean adults had a syllable-based strategy when segmenting Korean stimuli and French stimuli, as observed in previous studies on French speakers with French stimuli (e.g. Cutler et al. 1986, 1992). Murty et al. (2007) and Kim et al. (2008) argued that their studies extended the finding that speakers of genetically related, rhythmically similar languages showed similar segmentation

strategies (e.g. Sebastián-Gallés et al. 1992: Spanish, Catalan and French; Vrooman et al. 1996: English and Dutch), as their participants had no knowledge of the other language, which was rhythmically similar but genetically unrelated to their native language. Murty et al. (2007) and Kim et al. (2008) concluded that such perceptual tests of segmentation strategy, less labour-intensive than production experiments, were valuable in establishing languages' rhythm type, which apparently correlated with unit of segmentation strategy (e.g. French, syllable; Japanese, mora; Telugu, no clear preference for a particular unit, so mixed/unclassifiable rhythm type). From similar experiments on infants, Nazzi et al. (2006) reported evidence for the 'prosodic bootstrapping' proposal (Nazzi et al. 1998) that newborns are sensitive to rhythm type, using this to determine which segmentation strategy they develop according to ambient language: syllables for French (Nazzi et al. 2006), stress patterns for English (Echols et al. 1997).

In experiments which used segmentally degraded stimuli, Ramus et al. (2003) and White et al. (2007) found that French and English listeners could discriminate languages which were rhythmically distinct when measured with some rhythm metrics. Ramus et al. (2003) argued that such perceptual experiments were needed to assess the psychological reality of rhythm types. White et al. (2007: 1012) argued that 'perceptual validation of rhythm metrics is necessary', and that their results supported the use of these metrics 'in classifying perceptually salient aspects of speech rhythm.' Ramus et al. (2000) and Tincoff et al. (2005) found that primates (cotton-top tamarins) could also discriminate languages of different rhythmic types when hearing low-pass filtered speech. Tincoff et al. (2005: 33) interpreted this as evidence that humans and tamarins possess the same mechanism for language discrimination by rhythm, and 'that the [human] language acquisition device has recruited a general perceptual ability of the primate auditory system for language-internal purposes.'

1.2.5.3 Other disciplines

Some computational models of speech focussed on rhythm. Zellner Keller (2002) highlighted the importance of accurate rhythm in synthesising natural-sounding speech, and viewed rhythm as 'a multi-dimensional and non-linear cognitive construction' that primarily results from temporal organisation. Her (and colleagues') model, based on French and German, accounted for this multi-dimensionality and non-linearity by simulating rhythm through a temporal structure with harmonisation of segmental, syllabic and phrasal layers. Similarly, Barbosa (2002) argued that a deeper understanding of rhythm types would involve modelling several prosodic levels, and that segment-based rhythm metrics were too simplistic (cf. Nolan and Asu 2009). Barbosa (2002) computationally implemented a coupled-oscillator system for syllables and stress groups, by modifying for rhythm production an algorithm similar to McAuley's (1995) Entrainment Model for rhythm perception, which was instrumental in Port and Cummins' (1995, 1998) work. Later, Barbosa (2007) expanded his model to include several

levels of dynamic coupling between linguistic (syntactic and lexical) and acoustic sub-systems. Other computational models included: McLennan and Hockema's (2002) and McLennan's (2005), who aimed to provide evidence for Ramus et al.'s (1999) theory that %V, ΔC etc. are correlates of perceived rhythm, and integrate this theory with research on word-segmentation strategy; Villing et al.'s (2003), who wanted to implement automatic determination of P-centres.

Rhythm continued to feature in phonological research. Like Auer (1993), Schiering (2006, 2007) surveyed phonological descriptions of twenty genetically unrelated languages, and qualitatively scored them for ten stress-, syllable- and mora-related properties (cf. Dauer 1987). Schiering (2007) concluded, like Dauer (1983) and Grabe and Low (2002), that cross-linguistic rhythmic differences are gradient along a 'stress cline', and that rhythm type results from phonetic and phonological factors; he disagreed with Auer's (1993) proposal that a languages' overall rhythmic organisation characterises its entire phonological system. Schlüter (2005) proposed 'Rhythmic Grammar' based entirely on the 'Principle of Rhythmical Alternation' from Metrical Phonology. From an empirical investigation of English, Schlüter (2005) argued that this principle accounted for the distribution of so many morphological and syntactic phenomena, that a comprehensive English grammar must give it substantial presence.

Magne et al. (2004) called for interdisciplinary research between phoneticians and musicologists, to gain a better understanding of speech and musical rhythm. From an experiment investigating whether the cognitive processes involved in speech and musical rhythm perception are domain-specific or general, Magne et al. (2004) concluded that musical rhythm may be obligatorily processed, whereas speech rhythm processing may be modulated by attention, but both are processed on-line.

Some neurolinguistic research concerned rhythm, such as the following studies. Scott et al. (2006) investigated an English speaker with Foreign Accent Syndrome (FAS), who had a lesion near the primary motor cortex, and difficulty producing normal segmental and prosodic timing. Scott et al. (2006) concluded that FAS is a problem concerning prosody production, including intonation, timing and prominence, and that this arrhythmic speech was probably associated with cerebral perturbation to motor control. Ghitza and Greenberg (2009) presented to listeners synthesised utterances with inter-syllabic silences of various lengths. The waveform-energy fluctuations of the most intelligible stimuli were similar to the syllabic rhythm of natural (English) speech and matched the (theta) frequency range of known internal neural oscillators. Ghitza and Greenberg (2009) concluded that certain brain rhythms could be key to speech decoding, which is optimal when the utterances heard have a matching rhythm, and they called for brain-imaging studies with similar stimuli to elucidate the internal physiological processes. Jomori and Hoshiyama (2009) did just that, by measuring auditory event-related evoked potentials (AERPs, from electroencephalograms) when listeners responded to Japanese stimuli

that were temporally disrupted with inter-syllabic silences of various lengths (0-400ms). One AERP component (early negativity, EN) differed from normal when the silence in stimuli was at least 100ms, and other AERP components (N100-P150) differed from normal when the silence reached 400ms. Jomori and Hoshiyama (2009: 192) suggested that listeners might expect a syllable to occur at a certain time in rhythmic speech, but when it is delayed, their brain response reflects the fact that they noticed. These neurolinguistic studies provided insight into the brain mechanisms underlying rhythm production and perception, but more research is needed.

1.2.5.4 Summary (twenty-first century)

In this era, experimental phonetic research on rhythm was dominated by the development of a paradigm which quantified cross-linguistic differences in rhythm production. It was unclear whether researchers sought a categorical rhythmic typology or a more general systematic description of cross-linguistic rhythmic variation (Cummins 2002). Relatively few phonetic experiments concerned rhythm perception. Rhythm was also considered within computational modelling of speech and neurolinguistics. Most importantly, this era is the context for the present thesis.

1.2.6 Current state of experimental phonetic research on rhythm

This review has traced the development of experimental phonetic research on rhythm to its current state. The concept that language has rhythm is ancient; early scholars lacked the equipment to ground their ideas in anything other than auditory observation. As mechanical apparatus developed, experiments began to investigate these auditory impressions. Perceptual tests examined timing and prominence, whilst durational properties were measured in speech production. Such experiments appeared first in psychology, and later in phonetics, particularly as speech-recording equipment advanced. Following the proposal of a ‘stress-timing/syllable-timing’ rhythm typology in phonetics, timing and duration became the focus of phonetic experiments on rhythm, and great effort was put into testing this typology and proposing alternative theories. Most recently this has led to one dominating paradigm, which quantifies rhythm with statistical metrics; increasingly powerful and widespread computer technology has not hindered this. Despite these metrics’ domination, they have received criticism. Chapter 6 details the issues, which are summarised here as: the almost exceptionless preoccupation with duration; the need for perceptual validation of the metrics; and limited corpora, as most studies measured read (not spontaneous) speech, often short sentences, from only a few speakers.

A technological progression has allowed experiments to replace auditory observation of rhythm, but these have not provided clear results from which to form a consensual model of rhythm. At the Empirical Approaches to Speech Rhythm workshop (UCL, March 2008), there was talk of an impasse and a need for new directions in rhythm research, e.g. ‘to nudge the PVI

out of the rut' (Nolan and Asu 2009: 68). Where can we now look to tackle the challenge of understanding the complexity of rhythm? §1.2 stated that current experimental phonetic research on rhythm generally does not:

- (1) empirically investigate perception (from a non-typological perspective),
- (2) measure acoustic cues other than duration,
- (3) test whether native language/dialect influences speakers' perception of rhythm.

Arguably, these gaps in research are partly due to the fact that the field developed in response to claims for (1) a rhythm typology (2) based on timing (3) grounded in English speakers' auditory intuitions. 'Stress-timing/syllable-timing' left a profound trace; these terms still pervade rhythm literature and textbooks (e.g. Laver 1994: phonetics, Fox 2000: phonology), even though authors may no longer use the terms to mean that a categorical typology exists. Some perceptual experiments on rhythm have recently appeared in phonetics, but most had a typological perspective. Kim et al. (2008) and Murty et al. (2007) argued that listeners' preferred word-segmentation strategy was a means of identifying their native language's rhythm type. Ramus et al.'s (2003) and White et al.'s (2007) experiments tested whether listeners could discriminate/categorise languages that were reportedly of different rhythm types; Ramus et al. (2003) viewed rhythm typology as categorical, though probably involving more types than stress-/syllable-/mora-timing, whereas White et al. (2007) viewed it as non-categorical. As Pamies Bertrán (1999) and Nolan and Asu (2009) suggested, it would be better science to keep an open mind about the nature of speech rhythm, rather than limit investigation to typology. Filling the three gaps outlined above would benefit from an imaginary rewind to pre-typological times.

There are exceptions to the three generalisations above. After work began on this thesis, Kohler (2009a) proposed a 'new paradigm of rhythm research' with three parts: investigating various acoustic cues' contribution to perceived prominence and hence rhythm; perceptually evaluating the rhythmicity of different speakers' speech production; applying the same experimental designs to various languages with different rhythmical structures. Some experiments were in preparation (according to Kohler 2009a), but only a contribution to the first part of the new paradigm has so far been reported (by Kohler 2008). Disyllabic /baba/ utterances with systematic duration, f_0 and overall acoustic energy manipulations were presented to German listeners, who indicated whether the first or second syllable was more prominent. Duration and acoustic energy were less effective cues to prominence than f_0 was, though Kohler (2008) stated that listeners from other language backgrounds should be tested, because acoustic cues probably combine differently to create perceptual prominence patterns in different languages. Barry et al. (2009) and Niebuhr (2009), discussed in chapter 5, have also recently conducted phonetic non-typological research on rhythm perception, including cues other than duration and listeners of

different languages. The following sections present evidence from phonetics and other disciplines which demonstrates that the three gaps in research outlined above are worth pursuing.

1.2.6.1 Empirical investigation of perception

Evidence that we should investigate listeners' rhythm perception comes from the fact that rhythm research emerged from auditory impressions. Many researchers over several centuries, including throughout the controversy over rhythm typology, have noted one unequivocal fact: languages *sound* (i.e. are *perceived* as) rhythmically different (e.g. Classe 1939, Dauer 1983, Fletcher 1991, Roach 1982, Sievers 1912, Steele 1775 etc.). The earliest psychology experiments on rhythm were perceptual, but as speech recording instruments developed, so did a preference for supposedly more accurate and 'scientific' production research (see e.g. Brown 1908: 13, Wallin 1901: 70-1). However, since the development of technology allowing more complex psycholinguistic experimental paradigms and brain-imaging techniques, investigating perception cannot be seen as less worthy.

1.2.6.2 Measurement of acoustic cues other than duration

Evidence that rhythm involves non-durational acoustic cues comes from some phonetic (and psychology) experiments on rhythm, and some on prominence. Low (1998) demonstrated that rhythm could be quantified just as effectively with PVIs of amplitude and spectral dispersion as with durational PVIs. Steiner (2005) concluded, after reviewing his and others' data on rhythm metrics, that prominence-lending cues other than duration were needed to capture cross-linguistic rhythmic differences successfully. P-centre research suggested that amplitude envelope (and duration) may influence perceived isochrony in syllable sequences (e.g. Howell 1988, Pompino-Marschall 1989, Scott 1998). Lee and Todd (2004: 243) predicted that any prominence-lending cue could be used to distinguish languages' rhythms; their data from rhythm metrics incorporating f_0 , intensity and duration were insufficient to confirm this, but they called for further research on the interaction of multiple cues to rhythm. Tilsen and Johnson (2008) believed that their amplitude-envelope-based rhythm metric could improve duration-based rhythm research.

Psychological research has considered that both prominence and timing are involved in perceived rhythm (see e.g. Fraisse 1982). Likewise in phonetics, prominence and timing have long pervaded discussion on rhythm (see e.g. Kohler 2009a). Recently though, experimental phonetics has focussed on timing. In all discussion of 'stress-timing', few rhythm researchers examined the various acoustic cues to 'stress' (i.e. prominence), and instead viewed it as an abstract phonological concept, as in e.g. *Metrical Phonology*. This is unsurprising given the confusion surrounding this term 'stress'. Essentially, it describes a speech/linguistic unit (e.g. syllable) that perceptually 'stands out' compared to surrounding units, but various terms with various meanings

have been interchangeably used, e.g. ‘accent’, ‘accentuation’, ‘emphasis’, ‘force’, ‘intensity’, ‘prominence’, ‘stress’ (see e.g. Kohler 2009a: 30, Fox 2000: 114-5). Often these terms refer to ‘lexical stress’ (an abstract word-level phonological property in some languages), or to the speaker’s singling out of words for pragmatic function; they can also more generally mean the perceptual salience of any syllable with certain acoustic properties relative to its surroundings (Kohler 2008: 258). In this thesis, ‘prominence’ is used, since it is the most neutral and relevant term for this research (see chapter 2).

If we assume that rhythm involves prominence, but look beyond abstractness, the non-primacy of duration is evident from many phonetic experiments which identified various acoustic cues to prominence. Fox (2000: 120-22), Lehiste (1970: 106-41) and Crystal (1969: 113-20) give overviews of prominence production experiments. However, the presence of an acoustic feature in recordings does not necessarily prove its perceptual significance. Therefore, the following perceptual experiments are more relevant to perceived rhythm. Fry (1955, 1958, 1965) investigated prominence perception with synthesised English words like *object* and *digest*, understood as nouns or verbs if lexical stress is realised on the first or second syllable respectively. Various acoustic properties were manipulated in stimuli, and listeners’ responses indicated that increased duration and higher f₀ were more important prominence cues than increased intensity and vowel formant structure. Morton and Jassem (1965) adapted Fry’s experimental design for various languages, by synthesising and manipulating nonsense disyllables, so listeners judged according to general salience rather than lexical stress realisation. For English listeners, raised f₀ primarily determined prominence; increased intensity and (contrary to Fry 1958) duration were not effective prominence cues. For Polish listeners, raised f₀ was also the most effective cue, and increased duration was more effective than for English listeners (Jassem et al. 1968). Experiments similar to these were conducted for several languages e.g.: Rigault (1962) (French); Isačenko and Schädlich (1966), Gutknecht (1972), Kohler (2008) (German); Gussenhoven and Blom (1978) (Dutch); Bertinetto (1980) (Italian); Sautermeister and Eklund (1997) (Swedish); Llisterra et al. (2003) (Spanish); see also studies cited by Cutler (2005: 270). The overall take-home message from these experiments is that variation in duration, f₀, intensity and spectral quality can *all* potentially cue prominence (though little perceptual research has concerned the latter; see e.g. Sluijter et al. 1997), and each cue’s significance varies cross-linguistically. Most rhythm researchers in experimental phonetics have apparently lost sight of these findings from the same field. Furthermore, a few experiments, not specifically investigating prominence, provided evidence that two cues, f₀ and duration, *interact* in perception (e.g. Lehiste 1976), though others did not replicate this (e.g. Rosen 1977a, 1977b) (see chapter 3). One explanation for these conflicting results could be that listeners’ native language differed between studies.

1.2.6.3 Investigation of native-language influence

Evidence that native language might affect rhythm perception comes from conjectural comments by rhythm researchers, and empirical research in (mainly segmental) phonetics. According to Fletcher (1991) and Wenk and Wioland (1982), Anglophone linguists had defined syllable-timing negatively (rhythm not sounding like theirs). Roach (1982) and Dauer (1983) suggested that these Anglophones possibly perceived e.g. French rhythm whilst influenced by their expectations from English prosody, which might account for linguistically diverse phoneticians' disagreements over rhythm types. To native French speaker Vaissière (1991b: 109), French rhythm was 'much more obvious than the rhythm in English, where the stream of units of information is thrown into disorder by the intrusion of a recurring strong stress. [...] The language-specific rules of French may indeed not be obvious to non-natives'. Researchers in other fields suggested that listeners' native language possibly influenced their responses in non-speech rhythm experiments, e.g. Fraise (1984: 25) (psychology) and Patel (2007: 173) (musicology).

Many experiments have demonstrated that native language influences perception of segmental phonological contrasts (see several examples in: Hume and Johnson 2001, Strange 1995). The classic example is Lisker and Abramson's (1970) cross-linguistic perceptual study of voicing distinctions. Others include Flege and colleagues' experiments (e.g. Flege 1993, Flege et al. 1999). Since native language influences segment perception, it might influence prosody perception, though some phoneticians/phonologists might argue that segment and prosody perception are quite different. The widespread interest in cross-linguistic variation in perception has generally not included experiments on rhythm (cf. Beddor and Gottfried 1995). This is perhaps partly linked to the widespread acceptance since early psychology experiments that perceptual grouping of non-speech sounds was governed by innate universal cognitive principles (Iversen et al. 2009). However, the following three studies were interested in Jakobson et al.'s (1952) claim that a series of knocks with every third being louder is perceived in different groupings by Czech, French and Polish speakers, influenced by their native language's word-/phrasal-prominence location. Iversen et al. (2009) used stimuli which were tones in recurring sequences of alternating soft-loud (amplitude) or short-long (duration). English and Japanese listeners perceived the duration-alternating tones as respectively short-long and long-short groups; no cross-linguistic difference in grouping occurred for amplitude-alternating tones. Iversen et al. (2009: 2268) interpreted this as evidence for native-language influence: short function words precede longer content words in English, but follow in Japanese. In an experiment with similar stimuli, Hay and Diehl (2007) found no difference in grouping behaviour between English and French listeners, and concluded that 'the perception of linguistic rhythm relies largely on general auditory mechanisms.' Bailey et al. (1999) tested English and Portuguese speakers' ability to learn patterns comprising five tones with systematically varied duration and f_0

representing different rhythms. A familiarisation session preceded a recognition task testing whether speakers generalised their knowledge from the familiarisation stimuli to novel stimuli. The results suggested that speakers exploited their native-language knowledge when learning new rhythms, since English speakers were significantly more biased against final stress than Portuguese speakers. Further evidence for native-language influence on perception might be interpreted from rhythm studies which did not highlight the cross-linguistic differences they observed. Miller's (1984) results demonstrated that English and French listeners perceived rhythm differently, though the methodology was problematic. Scott et al.'s (1985) and Benguerel and D'Arcy's (1986) results showed small differences between English, French and (in the latter study) Japanese listeners' responses to stimuli testing perceived isochrony.

For other prosodic perceptual phenomena, native-language influence has received more attention. Ambient/native language apparently affects infant/adult listeners' preference for a word-segmentation strategy based at a particular prosodic level, which some researchers have argued could be determined by the ambient/native language's rhythm type (e.g. Cutler et al. 1992, Murty et al. 2007, Nazzi et al. 2006). Dupoux et al. (1997, 2001, 2008, 2010) and Peperkamp and Dupoux (2002) presented evidence that prominence perception depends on listeners' native-language prosodic phonology. French, Finnish and Hungarian, but not Spanish, speakers exhibited 'stress deafness': they found it hard to distinguish or recall minimal pairs of disyllabic nonsense words with first versus second syllable prominence. Dupoux et al. (2008) suggested that speakers of French, Finnish or Hungarian, which lack phonologically contrastive stress, lack the cognitive mechanism to phonologically represent contrastive stress that Spanish speakers have.

As discussed above, several studies similar to Fry's (1958 etc.) together showed that the relative significance of different acoustic cues to prominence (f_0 , duration, intensity, spectral quality) varies cross-linguistically. Logically, if perceived rhythm is induced by a pattern of prominences and non-prominences in speech, it should depend on which acoustic properties the listener hears as prominent according to their native language. Suppose that a speaker of language X perceives prominent syllables primarily as those with increased duration, whereas a speaker of language Y perceives prominent syllables primarily as those with a distinctly dynamic f_0 . Speaker-X perceives rhythm-X according to how (relatively) longer syllables occur in a pattern; speaker-Y perceives rhythm-Y according to how syllables with dynamic f_0 occur in a pattern. This simplification ignores an interaction of differentially weighted cues, but illustrates that perceived rhythm might differ cross-linguistically due to the different relative importance of prominence cues (cf. Lee and Todd 2004, Steiner 2005).

Despite recent calls to further investigate whether rhythm perception depends on native language (e.g. Arvaniti 2009, Frota et al. 2002, Kohler 2009a), this is still a gaping hole in

research. In testing listeners of one language, we cannot generalise their results to human language, or deduce that rhythm perception is universally identical (cf. Logan and Pruitt 1995). Furthermore, rhythm research in general has involved a relatively small number of mainly Indo-European languages. Exceptions include Keane (2006) (Tamil), Murty et al. (2007) (Telugu), Auer (1993) and Schiering (2006, 2007) (several languages from different language families). ‘Standard’ languages have been the focus, with relatively little known about rhythmic differences between non-standard varieties. Exceptions include Low (1994, 1998) (Singapore and British English), and Frota et al. (2002) (Brazilian and European Portuguese); arguably though, Singapore English and Brazilian Portuguese have become recognised ‘standards’ in those countries, and dialectal variation may exist within them.

1.3 Aims and research questions

It may now seem obvious that non-durational cues and native-language influence in rhythm perception should be researched. According to Tilsen and Johnson (2008: 34), from a naïve perspective the exclusion of research on non-durational cues might seem odd, ‘but it is so common that it is almost never explicitly noted in methodological appraisals.’ Experimental phonetic research on rhythm now needs to consider the evidence in §§1.2.6.1–3, from phonetics and other disciplines, to develop new lines of research. From this need emerged the present study’s aims and research questions. Primarily, this thesis aims to investigate speech rhythm from three perspectives, which current experimental phonetic research generally ignores, and so for which data is lacking. These perspectives are:

- **A focus on perception.** The aim is to investigate perceived rhythm, to link the findings to the PVI, and make this metric perceptually informed.
- **The inclusion of f_0 .** The aim is to observe not just duration, but how f_0 and duration are interdependent in rhythm perception and production.
- **A cross-linguistic study.** The aim is to provide evidence for or against native-language/-language-variety influence on perceived rhythm.

From these aims, the following research questions were formulated:

- Does f_0 play a significant role in speech-rhythm perception, and how does its significance compare with that of duration? Are tonal and durational cues interdependent?
- Are native speakers of different languages (and language-varieties) sensitive to durational and tonal rhythm cues to different extents? If so, is rhythm perception more language-(-/variety-)specific than universal?

To answer these questions, a series of three perceptual experiments are reported, starting with a psychoacoustic task, leading to a more psycholinguistic task, which together form the basis for a task concerning perceived rhythm in sentences, the results of which then link to a fourth production experiment. More specific research questions underlie each experiment, and are stated at the beginning of the relevant chapter. The same experiments are conducted with Swiss German, Swiss French and French (i.e. from France) speakers. This allows between-language and between-variety comparisons: Swiss German versus two varieties of French; Swiss French versus French from France. Although French and German are frequently-cited examples of a ‘syllable-timed’ and ‘stress-timed’ language respectively, this thesis does not assume that a rhythm typology exists. These languages were chosen because they exemplify languages which simply sound rhythmically distinct (amongst other reasons detailed in chapter 2).

It would be desirable to include many more languages (cf. Beddor and Gottfried 1995: 226), and investigate other acoustic cues, e.g. intensity, vowel quality, spectral balance (see Sluijter and van Heuven 1996, Sluijter et al. 1997). However, the scope of this thesis only allows for detailed investigation of two cues and two languages, though two varieties of one. It is hoped that this research will encourage others to conduct similar experiments with other acoustic cues and languages. If f_0 plays a significant role in rhythm, future research on rhythm should focus more attention on it (and other acoustic cues) than is currently the case. If rhythm perception differs between languages/varieties, future research should question the validity of universal rhythm metrics and a ‘one-size-fits-all’ approach to investigating rhythm.

1.4 Outline of thesis

This chapter has set out the research questions after reviewing previous studies and identifying certain gaps in research that need filling. Chapter 2 details the languages/varieties under investigation, including their status in the countries where spoken and their prosodic characteristics relevant to the experiments. Chapters 3 to 6 each report one of four experiments, and are structured similarly: first a context for the experiment and its hypotheses are given, followed by the method, then the results, which are interpreted and discussed, and interim conclusions are drawn. Chapter 7 pulls previous discussions and conclusions together, leading to overall conclusions, implications and suggestions for future research.

The languages investigated: Swiss German; Swiss French; French

2.1 *Why Switzerland?*

The main purpose of this chapter is to detail the prosodic structures of Swiss German, Swiss French and French (henceforth SG, SFr, Fr), which will be relevant in later chapters. First, the reasons for investigating these languages are expanded from chapter 1, followed by a brief outline of the languages' historic and current social context. 'Why Switzerland?' is the title of Steinberg's (1996) book about the country's unusual political, social and linguistic situation. For the present thesis, Switzerland was chosen as the main fieldwork location (as well as Cambridge for Fr speakers) because two of its national languages, SG and SFr, were particularly appropriate for several reasons.

First, (S)Fr (i.e. SFr and Fr) and SG have different prosodic structures, which result in different tonal and durational properties when phonetically realised, and most important here are these languages' different-sounding rhythms (see §2.3). As chapter 1 proposed, perceived rhythm results from the acoustic multi-dimensionality of the speech signal, and might depend on listeners' native-language prosody. To test this we need to investigate languages whose prosodic structures are manifested with duration and f_0 (and potentially other properties) in divergent ways.

Second, rhythm research has generally not investigated non-standardised oral dialects; rhythm concerns what we hear and speak, i.e. oral language, so researching such dialects, like SG, seems equally as relevant to rhythm theories as researching standardised languages. Most rhythm studies have investigated speakers (of standardised languages) whose accent is generally accepted within a country as the prestigious spoken form of language associated with the official written form. For example, White and Mattys' (2007a: 505) speakers 'had accents of their native language that were not markedly different from the commonly accepted standard (i.e. Algemeen Nederlands, standard southern British English, français neutre, castellano).' Other authors give no such detail (e.g. Grabe and Low 2002), or include speakers' regional origin without comment (e.g. Dauer 1983). A few studies have found that two language varieties, each an accepted standard in different countries, are rhythmically distinct, e.g. Singapore/British English (Low 1998), Mexican/Castilian Spanish (Nolan and Asu 2009). This suggests that rhythm research needs to compare varieties of languages spoken in different countries/regions, like SFr and the Fr spoken in France (cf. Barry et al. 2003, Grabe 2002). Interestingly, Frota et al. (2002) found that although Brazilian and European Portuguese rhythms were distinct when quantified with rhythm metrics, listeners could only distinguish them when the segmentally degraded stimuli had

durational and tonal cues, not durational cues alone, which links to this thesis' investigation of both cues.

Finally, a practical reason for choosing SG and (S)Fr is that (standard) German and Fr are the two non-native languages that I can write and speak fluently. The experiments in this thesis were conducted over three three-week periods in Switzerland for SG and SFr speakers, and in Cambridge (UK) for Fr speakers (my permanent location as a doctoral researcher). In both locations, particularly Switzerland, it was crucial that I could conduct fieldwork (writing instructions, recruiting subjects, and communicating with them during the experiment) efficiently and without relying heavily on others' assistance. In SG-speaking Switzerland, it is common to hear standard German spoken at university (see §2.2.1), which is where the experiments were conducted.

2.2 Social context

The following sections (§§2.2.1–2), mainly drawn from Galloway (2007, the present author), place the languages in the context of the communities which speak them, to explain their current nature. Although Switzerland is multilingual in that it has four national languages, the country comprises four practically monolingual geographically separate speech communities (Werlen 2002) (see Figure 2-1). That means the majority of Swiss people grow up acquiring one native language, which depends on their region, and have little contact with the other languages until second-language classes in formal education (see Ambühl et al. 2003, Schläpfer 1982, Steinberg 1996).

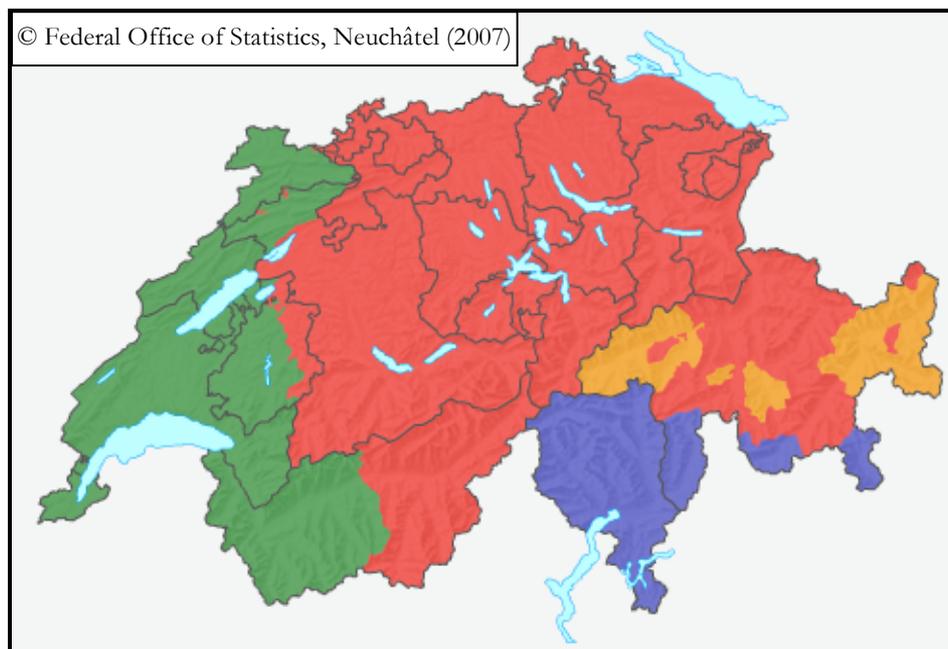


Figure 2-1 – Map of Switzerland showing data from the federal census in 2000, according to which the following percentages of Swiss nationals said their main language was: **(Swiss) German, 72.5%**; **French, 21.0%**; **Italian, 4.3%**; **Romansch, 0.6%** (Haug 2002)

2.2.1 Swiss German

SG (*Schwyzerdütsch*) is the collective name for the linguistically heterogeneous Alemannic dialects spoken since centuries ago within Switzerland (red area in Figure 2-1). These differ so much from (standard) German, that many are not mutually intelligible with it (Haas 1982). Each dialect has distinctive phonological, syntactic and lexical characteristics differentiating it from other SG dialects (detailed in the *Sprachatlas der Deutschen Schweiz* 1962-2003). Neighbouring dialects are mutually intelligible along a dialect continuum, and those in the far-east of SG-speaking Switzerland are generally very different from far-western ones.

Traditionally, a SG speaker refers to his/her dialect by the canton, e.g. Bern German, Zürich German etc. (Haas 1982), but even within cantons, variation exists from town to countryside, and between towns. Generally Swiss Germans (henceforth SGs) intuitively know where in their local area a person is from by hearing them speak (Lievano and Egger 2005: 4-5, Werlen 2000: 143). However, some SGs have, as one experiment participant remarked, a ‘mish-mash’ of dialects from having lived in various cantons/towns. This increasing social mobility is leading to some homogenisation of SG dialects, within cantons, e.g. Bern (Werlen 2000), and nationally. According to Reese (2007), the centre of this national process is Zürich, whose dialect has more pan-Swiss characteristics (and influence from standard German) currently than centuries ago, so Zürich residents can to some extent justify saying that they speak ‘Swiss’ German. Beilstein-Schauvelberger’s (2007) textbook entitled *Züritüütsch: Schweizerdeutsch* (‘Zürich German: Swiss German’), widely used by standard German speakers learning SG, also illustrates Zürich German’s status. Nevertheless, according to Reese (2007: 5) citing Christen (1998), ‘[p]robably, there will never be a uniform Swiss German’. For this thesis, Zürich was chosen for fieldwork partly thanks to the offer of a phonetics laboratory to work in, and partly because it offered a pool of participants whose native language was maximally homogenous given the traditional heterogeneity of SG. To recruit a group of linguistically homogenous participants is difficult enough when investigating ‘standard’ languages (Beddor and Gottfried 1995), but it is even harder when the investigated ‘language’ comprises non-standardised dialects.

Haugen’s (1966) often-cited model of language standardisation identifies the following stages: *Selection* of one dialect (norm); *Codification* of the norm into official dictionaries and prescriptive grammars; *Elaboration* of the norm’s function in the community; *Acceptance* by the community, i.e. the norm becomes a country’s standard language. Unlike several (European) countries, SG-speaking Switzerland has not experienced this standardisation process in oral language. No single dialect is the accepted norm, dictionaries are informal and grammars are de-not pre-scriptive, no official orthography exists, and the dialects’ function is limited mainly to spoken language. SG-speaking Switzerland is diglossic: ‘two different languages or language varieties are used in a single speech community, with each variety being largely reserved for certain purposes’ (Trask 1996: 69). SG is spoken in most everyday situations and acquired from

birth, whereas ‘Swiss standard German’ (SstG) (*Schweizer Hochdeutsch*) is the written language, taught in schools, and is spoken in international communication, academia, federal-level political discourse and many national media (Rash 1998: 52-70). However, recent communication methods, like text messaging, internet chat rooms and social-networking websites, are encouraging the use of written SG (Siebenhaar 2006). Between SstG and the standard German of Germany, some phonological, morphological, syntactic and lexical differences exist, which are generally minor compared to differences between SG and SstG (Haas 1982, Siebenhaar and Vögeli 1997), except that a Swiss accent is highly recognisable. Some speakers, who are rarely in situations which require them to speak SstG (e.g. university, national media), may feel they sound inferior to the Germans and unnatural or stilted when speaking SstG (Haas 1982).

2.2.2 (Swiss) French

France has experienced language standardisation, the following outline of which draws from Lodge (1993) and Walter (1988), using Haugen’s (1966) model. *Selection* of a norm occurred in the 1100s-1200s; of numerous Gallo-Romance dialects descended from Latin, the one spoken by the politically powerful in what we now call Paris, France emerged as more prestigious. *Elaboration* of the function of early Fr (the norm) proceeded, particularly in the 1400s-1500s, and Fr replaced Latin as the written language. *Codification* occurred in the 1500s-1700s, with the publication of official dictionaries and prescriptive grammars. *Acceptance* by the majority of France’s population only finished within the past century; gradually a geographically and socially more widespread community came to speak ‘standard’ Fr, replacing their local (non-Parisian) dialects. Contemporary Fr displays variation, mainly depending on speech style and speakers’ regional origin within France or social status (Lodge 1993: 230-1). Generally speakers with lower social status display more linguistic features influenced by the formerly spoken local dialects than speakers with (or striving to have) higher social status, who show little deviation from Parisian Fr, which is regarded as superior to regionalisms, linked to Paris’ dominance in politically, economically and culturally centralised France (Ball 1997: chapter 5, Lodge 1993: 235).

Like France, and unlike SG-speaking Switzerland, francophone Switzerland has experienced language standardisation. Across the blue area in Figure 2-1, various Francoprovençal dialects were once spoken, which had similarities to, and differences from, dialects once spoken in France (Knecht 1982, Miller 2007, Walter 1988). During the *Elaboration* and *Codification* of Fr in France, the language’s prestige spread to the Francoprovençal area, where the dialects were not very cohesive due to geographical, political and religious divisions, which led to the *Acceptance* of Fr in that area (Miller 2007). According to the 2000 Swiss census, only 1% of the population in francophone Switzerland said they speak a Francoprovençal dialect (at home) (Lüdi and Werlen 2005). The Fr that became accepted in Switzerland (SFr) has been described as distinct from the Fr of France. Often-cited examples of SFr characteristics are:

phonemic vowel-length contrasts in word-final syllables; use of *vouloir* (not *aller*) for forming the future tense; Germanisms in vocabulary, e.g. *poutser* (from *putzen*, ‘to clean’); retained use of archaic words e.g. *septante* (‘seventy’) and *nonante* (‘ninety’) (see Knecht 1982, Voillat 1971, Walter 1988). According to anecdotal data, local variation exists within SFr between towns and cantons which is minor relative to variation between SFr and Fr (Arès 1994, Manno 2004, both cited by Miller 2007, Knecht 1982, Singy 1996); some local variation might be due to influence from formerly spoken Francoprovençal dialects (Miller 2007).

However, Voillat (1971: 238) claimed that SFr was moribund, increasingly homogenised with Fr, because, like in France, those who wanted to climb the social ladder strove to speak prestigious Parisian Fr. Similarly, Offord (1990: 218) stated that SFr ‘is shedding many of its former distinctive traits and is becoming very similar to standard French. [...] In fact no peculiarities, apart from a few lexical items, are exclusive to [SFr]; they also occur in the French of Belgium or in the neighbouring dialects of France’. Walter (1988: 247-9) described the situation similarly, though neither author mentioned actual data. Yet empirical research demonstrates that SFr was and currently is distinguishable from Fr, mainly in phonology/phonetics and vocabulary, as the following studies exemplify. (Research on prosodic variation between SFr and Fr is detailed in §2.3.2.4.)

Métral (1977) and Walter (1982) found that SFr, unlike Fr, had phonological vowel-length contrasts in word-final syllables, e.g. *vit* /vi/, *vie* /vi:/; *roux* /ʁu/, *roue* /ʁu:/. Both studies had a perhaps dubious method – asking non-linguists: (1) Do you pronounce X and Y identically? (minimal pairs inserted); (2) If there is a difference, is it in length, or something else? In an auditory analysis of SFr corpus data, Andreassen (2006) found that various vowel-length contrasts in word-final closed syllables were less stable than those in word-final open syllables, and /e:/ in word-final open syllables tended to diphthongise to /ej/. In an acoustic-measurement experiment and a word-recognition experiment, Grosjean et al. (2007) found that SFr speakers, but not Parisian Fr speakers, produced and unproblematically perceived various vowel-length contrasts in word-final open syllables. In perceptual experiments involving priming or discrimination tasks, Dufour et al. (2007) and Brunellière et al. (2009) found that the /e/-/ɛ/ vowel-quality contrast, which is reportedly merging in (northern) Fr, was still perceptually real for SFr speakers, though neuroimaging data suggested that they found it harder to cognitively process than the non-merging /ø/-/y/ contrast. Woehrling and Boula de Mareüil (2006) found that Parisian Fr speakers could distinguish SFr easily from other regional Fr varieties, when they heard short recorded extracts of speech by (S)Fr speakers from various regions. Schoch (1978) concluded from her investigation of second person pronoun (*vous/tu*) usage by SFr speakers that SFr was a variety distinct from Fr. Bayard and Jolivet (1984) and Singy (1996) used sociolinguistic questionnaires to investigate attitudes of SFr speakers concerning their language. Bayard and

Jolivet (1984: 153) concluded that there was little feeling of linguistic insecurity amongst their subjects; half thought the ‘best’ Fr was spoken in France, and half thought Switzerland. These subjects then participated in further tasks, including evaluation of SFr and Fr accents for traits like friendliness, and a priming experiment with words with different meanings in SFr and Fr; the results suggested that SFr and Fr were distinct. Singy (1996: 258) concluded that linguistic insecurity *was* noticeable amongst his subjects, and that SFr speakers were at a linguistic periphery where the central reference point was Paris. Evidence for this included: four in five subjects attributed the ‘best’ Fr accent to France; one quarter said they did not like their Swiss accent; one third said they tried to hide their accent. These subjects identified several characteristics of SFr that still distinguished it from Fr: ‘accent’, ‘lexical regionalisms’ and ‘slow speed’ were the top three responses (87.6%, 47.0%, 25.4% of subjects respectively) (Singy 2001). The preface to the *Dictionnaire Suisse Romande* (2004), which documented lexical dialectology conducted by the *Centre de dialectologie et d’étude du français régional* (Neuchâtel University), stated that differences between SFr and Fr mainly concern vocabulary and phraseology.

2.2.3 Summary (social context)

The linguistic situation in francophone Switzerland is different from that in SG-speaking Switzerland. Knecht (1979: 249) called francophone Switzerland ‘a politically Swiss France or a linguistically French Switzerland’ (translation RC). SFr and Fr are different varieties (or, arguably, accents) of a standardised language each spoken in separate countries, but the dominant social/political prestige of Fr influences SFr, so Swiss linguistic individualism is reduced. In contrast, we might call SG-speaking Switzerland ‘a politically and linguistically Swiss Switzerland’. Non-standardised SG dialects are more than phonologically different from standard German, and no external dominant social prestige influences SG, so Swiss linguistic individualism is retained. SstG and standard German are relationally equivalent to SFr and Fr, though SstG is seen as a second (i.e. non-native) language by many SGs (Lievano and Egger 2005).

2.3 Prosody

Reference will be made back to the following descriptions of SG and (S)Fr prosody when the experiments of this thesis are reported (chapters 3-6). For each language, the description comprises three sections: rhythm, prominence, intonation. The first mainly concerns timing/duration, since this has been the research focus to date. As discussed in chapter 1, prominence is also highly relevant to rhythm. It will become clear why ‘prominence’ (not stress, accent(uation) etc.) is the most neutral term to use when comparing SG and (S)Fr. Recent intonation research (particularly in the Autosegmental-Metrical framework, as in Pierrehumbert 1980) demonstrates the nature of f_0 contours associated with prominent syllables in various

languages, so intonation data are also relevant in this thesis which concerns f_0 (and duration) in rhythm perception.

2.3.1 Swiss German

SG has been subject to relatively little experimental phonetic research. According to Siebenhaar (2005: 343), only recently have experiments (e.g. Auer et al. 2000) investigated the early impressionistic observations of regional prosodic variation in non-standard German dialects (e.g. Bremer 1893, Sievers 1912). Although twentieth-century dialectology extensively documented the SG dialects' segmental phonology (*Sprachatlas der Deutschen Schweiz* 1962-2003), data on SG prosody amount to a few conference presentations (e.g. Fitzpatrick-Cole 1999), and Siebenhaar and colleagues' papers reporting results from two Swiss research projects, which examined various interacting aspects of SG rhythm and intonation using speech synthesis technology (e.g. Häsler et al. 2005, Siebenhaar et al. 2004). Limited descriptions of prosody also appear in recent overviews of SG (Fleischer and Schmid 2006, Reese 2007). All these sources of information on SG prosody were consulted for the following sections, which sometimes refer to standard German and SstG for comparison.

2.3.1.1 Rhythm

Germanic languages have generally been grouped as 'stress-timed' or 'stress-based' (e.g. Dauer 1983, Grabe and Low 2002, Ladefoged 2001, Ramus et al. 2003). According to Reese (2007: 11), in phonological terms SG 'can only be a stress-timed language' since it has trochaic feet (a prominent followed by a reduced syllable). Conversely, Nübling and Schrambke (2004) argued that SG has many traits of a 'syllable' language in Auer's (1993) phonological framework (which was based on prosodic units' phonological structure, with a continuum between prototypical 'syllable' and 'word' languages; see chapter 1). Nübling and Schrambke's (2004) examples included: SG lacks glottal stops, allowing resyllabification across word and morpheme boundaries, resulting in mostly open syllables (cf. Siebenhaar et al. 2004); in some dialects (bordering the francophone region) the acoustic contrast between prominent and non-prominent syllables is relatively 'weak', giving the impression of an accent more similar to Fr.

Phonetic experiments on rhythm in English, the canonical so-called 'stress-timed' language, outnumber those on other Germanic languages (see the number of references in Dauer 1983). A few studies using rhythm metrics (PVI and/or %V, ΔV , ΔC) found relatively high durational variability, of both vowels and consonants, for SG (Galloway 2007, Schmid 2001) and standard German (Barry et al. 2003, Grabe and Low 2002), which may reflect these languages' phonological structure. SG has phonemic consonant- and vowel-length contrasts and complex syllables; onsets and codas can contain large consonant clusters, which mainly result from historical processes whereby definite articles and prepositions, for example, have assimilated into

the following noun and lost their vowel, e.g. *dFrau* [pfrau] (the woman) *zBäärn* [tsbæ:rn] (in Bern) (compare standard German *die Frau*, *zu Bern*). SG has lexical stress, and in unstressed syllables vowels are reduced and consonant clusters are rarer (Fleischer and Schmid 2006, Reese 2007). These properties are consistent with the view that SG lies near the stress-based (Dauer 1983) or more stress-timed (Grabe and Low 2002) end of a rhythmic continuum.

2.3.1.2 Prominence

2.3.1.2.1 Phonology

Several general descriptions of SG have mentioned prominence from a phonological perspective (e.g. Stucki 1921, Keller 1961, Fleischer and Schmid 2006, Reese 2007). These noted that SG has lexical stress, with the initial syllable of polysyllabic words usually being prominent. Fleischer and Schmid (2006) explained that this results from stress being assigned to the lexical root, which is often the word-initial syllable, and a preference for initial lexical stress in loans and acronyms. Based on Maas' (1999) model for standard German, Reese's (2007) model of SG metrical structure proposed that feet are, in theory, disyllabic and trochaic, but deviations can occur, e.g. two successive reduced syllables (dactylic structure) or a prominent syllable is a foot by itself (a reduced foot). In words like *dFrau* [pfrau] and *zBäärn* [tsbæ:rn] (see §2.3.1.1), a disyllabic iambic foot has become a monosyllabic reduced foot through the assimilation of article/preposition and noun. Reese (2007) claimed that phonological-word-initial feet, often phrase-initial, are most prominent, but may be suppressed in a longer utterance if another word is in focus.

2.3.1.2.2 Phonetic correlates

To my knowledge, nobody has conducted perceptual research on cues to SG prominence which parallels the studies mentioned in chapter 1 (e.g. Fry 1955, 1958: English; Gutknecht 1972, Kohler 2008: German). We can examine production data, but the presence of an acoustic feature in the signal does not necessarily prove its perceptual significance. Siebenhaar and colleagues' research provides measurements of duration and f0 in (non-)prominent syllables. (Siebenhaar (2005) admitted that intensity and voice quality received less attention than timing and intonation in their research.) Reported in Siebenhaar et al. (2004) and Häsler et al. (2005), various acoustic measurements were taken from recorded interviews of three SG speakers. Vowels were significantly longer in prominent syllables (i.e. with lexical stress) than non-prominent syllables; consonant duration differed little between prominent and non-prominent syllables. Phrase-initial and phrase-final syllables were significantly longer than phrase-medial syllables. When all these duration data were applied in speech synthesis, the algorithm treated phonologically long and short vowels separately, because lengthening and prominence had been found to not change all vowel durations equally linearly (Siebenhaar 2005). To analyse f0

production, Häsler et al. (2005) used Fujisaki's (1984) modelling program, adapted for German by Mixdorff (1998), which divides the f_0 contour mathematically into local movements (related to words/feet, i.e. pitch-accents in Autosegmental-Metrical terminology) and utterance-global contours¹. Local movements occurred mainly on content words, on both prominent syllables (i.e. with lexical stress) and non-prominent syllables. (Siebenhaar and colleagues generally referred to prominence as *Akzent* 'accent', rather than *Betonung* 'stress'.) The speaker from Bern had a higher mean f_0 excursion on non-prominent than prominent syllables, and the speaker from Zürich the opposite. From similar analyses of durational and tonal properties of seventeen SG speakers' speech (Bern and Valais cantons), Leeman and Siebenhaar (2007) also found that: prominent vowels were longer than non-prominent ones; significant phrase-initial and phrase-final lengthening occurred; and f_0 movements occurred on both prominent and non-prominent syllables.

For standard German, according to Schneider and Möbius (2007), Jessen et al.'s (1995) production experiment found that increased duration was the most reliable correlate of prominence. Schneider and Möbius (2007) found that intensity was not a reliable correlate of prominence realisation in standard German, though spectral tilt (a measure of the balance between amplitudes in low-frequency and higher-frequency spectral domains) was, which replicated Classen et al.'s (1998) finding. Ulbrich (2006) found that SG speakers, when they spoke SstG, produced significantly longer vowels in prominent syllables than German speakers from Germany. Given that an individual's SG dialect influences their SstG pronunciation (Hove 2002), and that non-prominent vowels seem to be (durationally and spectrally) reduced more often in SG than in standard German (Siebenhaar and Vögeli 1997), duration might be a more significant cue to SG (than standard German) prominence. Confirmation of the perceptual significance of the cues identified from these production experiments is needed.

In sum, increased duration and f_0 movements seem important in the production of SG prominent syllables, though lengthening also occurs at phrase boundaries, and f_0 movements also occur on apparently non-prominent syllables. The following section discusses f_0 in more detail.

2.3.1.3 Intonation

The intonation of German dialects/varieties spoken in the south of German-speaking Europe is generally characterised by prominent syllables with a right-displaced f_0 peak, i.e. L*+H compared to H*+L in northern German (Gibbon 1998: 93), e.g. Swabian (southern Germany)

¹ Originally the program was designed for artificially producing f_0 in speech synthesis, but if reversed it can be used for analysing f_0 in natural speech (see Siebenhaar et al. 2004, Siebenhaar et al. 2006).

(Frey 1975, Kügler 2004), Tyrolean German (parts of Austria and Italy) (Barker 2005), SstG and SG (see below). According to Zimmermann (1998: 12), this bitonal displaced-peak prominence may explain why these southern varieties are described as ‘singing/musical’ (cf. Stedje 1999: 188), and sound strikingly different from standard (northern) German, even to linguistically untrained ears (cf. Häsler et al. 2005: 215). MacCarthy (1975: 26) illustrated the difference between standard German and SstG intonation as in Figure 2-2. Ulbrich’s (2002, 2004, 2006) corpora of standard German spoken by newsreaders in Germany, Austria and Switzerland support this illustration. Ulbrich (2002) also found that the Swiss produced nuclear accents with extremely late f₀ falls compared to speakers from Germany, and the Swiss speakers’ f₀, unlike the Germans’, remained high after the final pitch-accented syllable and fell sharply during the IP-final syllable. For Swiss speakers, but not Germans, Ulbrich (2006) found a strong positive correlation between vowel duration and f₀ excursion.

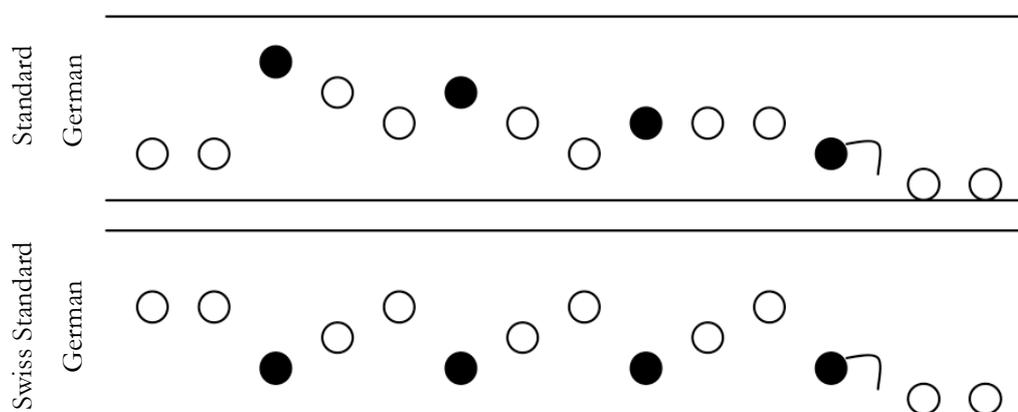


Figure 2-2 – Comparison of standard German and SstG intonation patterns

Hove (2002) demonstrated that SG speakers’ SstG pronunciation is influenced by their SG dialects; this is supported by the fact that the same intonation patterns have been observed in SG (Nolan and Hausmann 2005, Fitzpatrick-Cole 1999) as in SstG (Ulbrich 2002). Fitzpatrick-Cole (1999) (with speakers from the Bern canton) and Nolan and Hausmann (2005) (with speakers from various cantons) conducted Autosegmental-Metrical analyses of SG based on recordings of similar sentences. Fitzpatrick-Cole (1999) observed rises on prominent syllables, and suggested L*+H for the default pitch-accent; the +H peak often occurred long after the prominent syllable. A sharp fall occurred on IP-final words’ prominent syllable, which she analysed as an IP boundary tone (L%) that was optionally stress-seeking. Nolan and Hausman (2005) observed similar intonation patterns, but suggested (LH)*, not L*+H, for the default pitch-accent. Nolan and Hausman (2005) notated the sharp phrase-final falls, which (unlike pitch-accented syllables) were not perceived as prominent, with a HL phrase tone that docks onto the last unreduced (though not necessarily lexically stressed) vowel, where the low f₀ is realised.

Contrary to Autosegmental-Metrical transcription, Häsler et al. (2005) concluded that the realisation of pitch-accents is more gradient than categorical labels like L*+H suggest; in their data, f₀ rises sometimes started before prominent syllables' onset, sometimes after, and generally the longer the syllable, the greater the time interval was between the syllable onset and f₀ rise. Häsler et al. (2005) compared their SG data to some from standard German and SstG, and found that in SG and SstG, IP-final syllables had lower f₀ excursion on average than prominent (IP-medial) syllables, which had greater f₀ excursion on average than in standard German. Leeman and Siebenhaar (2007) replicated this finding for SG with several more speakers. Häsler et al. (2005) concluded that, consistent with Hirschfeld and Ulbrich's (2002) findings, intonation is marked at the stress-group/foot level in SG and SstG, and at a more utterance-global level in standard German, which presumably contributes to why they sound different. Likewise, with data from one speaker, Fleischer and Schmid (2006: 250) commented that SG 'seems to display larger overall F₀ range with a greater number of pitch movements' compared to standard German.

In sum, in SG f₀ rises are associated with prominent syllables, though the rise may start late and end within the syllable following the prominent one. A sharp fall often occurs in IP-final position, just before the boundary. Substantial f₀ movements may be more frequent and prominent at the foot level than the IP level.

2.3.1.4 Cross-dialectal variation

A few researchers working on SG prosody (e.g. Leeman and Siebenhaar 2007, Siebenhaar 2004) see cross-dialectal variation within SG as important to study. According to Häsler et al. (2005) and Leeman and Siebenhaar (2007), Zürich and Valais German have different-sounding prosody compared to Bern German. Both these studies found subtle cross-dialectal differences in durational properties, which could explain the audible cross-dialectal variation. Häsler et al. (2005: 209) admitted that differences they observed between their (just) two speakers' duration data could equally result from individual style rather than dialect, which was unlikely in Leeman and Siebenhaar's (2007) study with seventeen speakers. Siebenhaar et al. (2006) explained that if the timing and f₀ contour from a Bern German recording are resynthesised with the segmental signal from a Zürich German recording, or vice versa, the resulting utterances sound 'very unnatural'. These SG experts work in Switzerland, so have significant exposure to the dialects and are likely to hear subtle cross-dialectal differences.

Cross-dialectal variation was not investigated in this thesis, for which the fieldwork was conducted in Zürich (see §2.2.1 for reasons). In two experiments, only life-long residents of the Zürich canton participated, because if speakers had been included who had previously lived elsewhere, and so spoke a somewhat mixed dialect, this could have influenced their perception and production of Zürich German sentences (see chapters 5 and 6). In another two experiments, participants only had to be currently resident in Zürich, though most had lived there several

years. This more permissive criterion greatly facilitated participant recruitment; it was justified because the monosyllables used in the stimuli, unlike sentences, do not show cross-dialectal segmental variation, and were prosodically stylised, thus not representing the fine prosodic detail of any specific dialect (see chapters 3 and 4). For this thesis, ‘SG’ (rather than ‘Zürich German’) is appropriate because the primary concern is rhythm, and SG serves as an example language, though this presents the problem of finding a linguistically homogenous participant pool, a fairly good one of which Zürich happens to offer, given the heterogeneity of SG.

2.3.2 French

Investigation of Fr prosody amounts to a substantial body of research, comprehensively summarised in Lacheret-Dujour and Beaugendre (1999). The following description of Fr prosody largely concerns theories based on the Fr of France; §2.3.2.4 on SFr will identify some between-variety prosodic variation.

2.3.2.1 Rhythm

It has long been recognised that Fr rhythm sounds different from English rhythm (e.g. Abercrombie 1967, Jones 1956: first published in 1918, Steele 1775). Fr became the canonical so-called ‘syllable-timed’ language, though strict syllable isochrony has not been found from duration measurements of Fr (e.g. Roach 1982, Wenk and Wioland 1982), and alternative ways to model Fr rhythm have been proposed. According to Wenk and Wioland (1982), rhythmic groups (with durationally unequal syllables) are regulated by the group-final syllable in Fr and group-initial syllable in English, hence ‘trailer-timed’ and ‘leader-timed’ rhythm respectively. Fletcher (1991) commended their theory for proposing Fr rhythmic units larger than the syllable, which had become almost thoughtlessly accepted as the rhythmic unit in ‘non-stress-timed’ languages. From her own measurements of syllable and rhythmic-group durations, Fletcher (1991: 208) concluded that Fr ‘shares many timing patterns with languages of purportedly different rhythmic structure.’ According to Vaissière (1991a, 1991b), Fr has a multi-layered rhythm with three interacting perceptual units (breath group, prosodic word, CV syllable); in any utterance, depending on speech style, any of these groups can be made perceptually most emergent. Vaissière (1991b, citing Delattre 1966b) claimed that in Fr, the syllable is particularly salient because syllables tend to have similar characteristics, e.g. little vowel reduction, mostly CV structure, or tendency for resyllabification thus in connected speech. Vaissière (1983: 64, 1991b: 118) called Fr a ‘boundary language’ and English a ‘stress language’: in English, lexical stress competes with final lengthening to be prosodic groups’ most salient feature; in Fr, there is no lexical stress to compete with final lengthening, so prosodic-group boundaries are prominent. Syllables and stress may be more salient in Fr and English respectively, since in experiments on word-segmentation strategies, Fr and English listeners showed syllable-based and stress-based strategies respectively (Mehler 1981, Nazzi et al. 2006, Smith et al. 1989). Cutler et al. (1992)

proposed abstract structures to what speakers produce and perceive, so further research was needed on the phonetic realisation and perception of acoustic cues to Fr prominence. From his and others' observations, Di Cristo (2000) suggested that the realisation of underlying prominences depends on various factors including speech rate: generally, alternating initial and final prominences are produced, but fast speech may have fewer URs, as several words group into one prominence-unit, and initial prominences are inhibited; in careful/deliberate speech, initial prominences may be made more salient than usual, for emphasis. According to Di Cristo (2000), some complexities of Fr prominence structure may contribute to rhythmic variability in its phonetic realisation.

In sum, Fr does not have lexical stress; prominence always occurs prosodic-group finally, and optionally prosodic-group initially. Prosodic groups of roughly two to five syllables are the prominence domain. Several analyses of Fr prosody from speech production data have observed prominence patterns similar to the phonetic realisation described by Di Cristo (2000) (e.g. Jun and Fougeron 1995, 2000, Mertens 1987, 1992, 1993, Post 2000: these examples are discussed with intonation (§2.3.2.3), since they focussed on f₀ contours).

2.3.2.2.2 *Phonetic correlates*

Some production experiments have investigated Fr prominence, though these data do not necessarily prove cues' perceptual significance. From an oscillograph recording, Parmenter and Blanc (1933) measured pitch, intensity and duration, and concluded that higher pitch is the 'principal means' of Fr prominence, though the highest pitch occurred on the longest (often group-final) vowels, suggesting that duration was also significant. Conversely, from another production experiment, Delattre (1966b: 68) concluded that pitch is an important, but not indispensable, prominence correlate, and duration is highly significant. Both studies only recorded one speaker, and neither referred to different prominence types, unlike Benguerel (1971, 1973) who investigated 'emphatic' (group-initial) and 'unemphatic' (group-final) prominence. (He cited (1973: 21-22) twenty-seven terms previously used by other authors for Fr prominence, e.g. 'pitch-', 'intensity-', 'musical-', 'word-', 'group-', 'logical-', 'consonantal'-prominence.) In one experiment, which concerned unemphatic prominence, he measured vowel duration in six speakers' recordings, and concluded that increased duration in rhythmic-group-final syllables cues prominence when f₀ simultaneously falls, but when f₀ simultaneously rises, prominence is cued by the rise (Benguerel 1971). Another experiment elicited both prominence types; he measured air flow, sub-glottal pressure, f₀, intensity and duration in two speakers' recordings. For group-final prominence, neither air flow nor sub-glottal pressure were viable correlates, but increased duration was; for group-initial prominence, increase in sub-glottal pressure was the best physiological correlate, and f₀ the best acoustic correlate (Benguerel 1973). Crompton (1980) measured, in four speakers' recordings, the duration of 'stressed' and

‘unstressed’ syllables. He did not indicate how ‘stress’ was determined; as his analysis did not include group-final syllables (since he criticised experiments on English prominence like Fry’s (1958) and Lieberman’s (1960) for confounding prominence with phrase-final lengthening effects), he must have observed group-initial prominence. Crompton (1980) concluded that duration was not an important correlate of Fr prominence, though intensity might be, according to another (unreported) analysis he had done. Lacheret-Dujour and Beaugendre (1999: 120) summarised production data related to Di Cristo’s (2000) model, and concluded that duration, intensity and f_0 were significant markers of IP-final prominent syllables. Various studies not specifically concerned with prominence have shown that Fr has extensive group-final lengthening (Fletcher 1991, Kamiyama 2003, Vaissière 1991b).

In a perception experiment, Rigault (1962) asked Fr speakers to label the ‘most prominent’ syllable in synthesised disyllabic words with f_0 , duration and intensity manipulations; the most significant cue to prominence was found to be higher f_0 , though it is unclear whether speakers related their responses to group-initial or to group-final prominence (cf. Benguerel 1973, Morton and Jassem 1965). Mertens (1991) asked twenty phonetically untrained Fr speakers and one phonetician to indicate ‘stressed’ syllables in short recordings of Fr. Contrary to Fónagy’s (1980) finding of divergent reactions when Fr speakers were given this task, Mertens (1991: 220) concluded that the agreement amongst his speakers ‘indicates the perceptual reality of prominence’. Mertens (1991) ran a regression analysis on various tonal, durational and intensity-related measurements of syllables judged ‘stressed’; the best predictors of ‘stress’ were duration relative to preceding syllables, and syllable nucleus duration. The phonetician dichotomised his judgments as group-initial/-final; the f_0 of group-initial and group-final stressed syllables was comparable, whereas the duration of group-initial stressed and non-prominent syllables was similar, and shorter than group-final stressed syllables.

In sum, higher f_0 and increased duration may both be correlates of Fr prominence. Di Cristo (1999: 162, 2000: 40) claimed that limited data (Artésano et al. 1995, Lyche and Girard 1995) supported the idea advanced by ‘several authors’ that group-initial prominence is cued by f_0 variation, whereas the primary cue to group-final prominence is syllable lengthening, though f_0 is not insignificant (see also Vaissière 1991b: 115-16), but he admitted that further perceptual data were needed.

2.3.2.2.3 *Controversy over Fr prominence*

The following summary of the controversy over Fr prominence is based on Di Cristo’s (1999) comprehensive review. Many linguists, including Fr speakers, came to accept that Fr is ‘une langue sans accent’ (a language without stress). Arguments which led to this idea (listed by Di Cristo 1999: 157-62) include: the lack of lexical stress; the lack of secondary metrical stress; the refusal to count ‘emphatic’ (group-initial) prominence as actual prominence; the non-

independence of prominence from intonation, since f_0 movement also occurs on the prominent group-final syllable; the ‘weak’ realisation of prominence.

The lack of lexically contrastive prominence was observed by early twentieth-century Fr phoneticians (e.g. Fouché 1933-34, Grammont 1914). They noted that the final syllable of citation-form words or rhythmic/sense groups was always prominent, and various names were given to this, e.g. ‘final’, ‘tonic’, ‘rhythmic’ prominence (Di Cristo 1999: 163). Non-Fr linguists (e.g. Fletcher 1991, Halle and Vergnaud 1987, Hyman 1975) adopted this view of Fr prominence (Di Cristo 1999: 163). The early phoneticians had also observed word-/group-initial prominence, but believed it was regional, vulgar or harmful to ‘beautiful’ Fr (Di Cristo 1999: 163). Rossi’s (1980) experiment (entitled *Le Français, langue sans accent?*) investigated the interdependence of prominence and intonation, by analysing recordings of homophonous sentences that were only distinguishable by prosodic features. Rossi (1980) concluded that Fr is a language without stress, because stress and intonation are not distinct units, in form or function, and that intonation has a fundamental function in realising syntactic structure, so prosodic groups smaller than IPs are unlikely to exist. Several authors suggested that prominence is less perceptually striking in Fr than other languages (e.g. those cited by Di Cristo 1999: Beckman 1993, Dauer 1983, Fouché 1933-34, Hall 1946, Nyrop 1963, Tranel 1987). According to Wenk and Wioland (1982), it is unsurprising that Fr prominent syllables, marked by lengthening, slip the attention of Germanic-language speakers’ ears, which are tuned to large f_0 movements signalling prominence. Native Fr linguist Vaissière (1991a: 259) also talked about no ‘clear strong beat’ in rhythmic groups (cf. native Fr linguists cited above: Fouché 1933-34, Tranel 1987).

Di Cristo’s (1999) arguments against the ‘langue sans accent’ idea include: group-initial prominence should not be ignored just for sociolinguistic reasons; group-initial and group-final prominence are both prominence, but realised with different phonetic correlates (see §2.3.2.2.2); it is counter-intuitive to propose that a language does not have metrical structure, though it may be closely related to tonal structure; there is evidence for smaller prosodic groups within IPs, each with a final prominent syllable.

Fónagy (1949, 1980) contributed greatly to reversing the view that Fr prominence was as simple as once thought. Fónagy (1980: 71-77, 130-39) demonstrated that Fr prominence was not fixed and predictable, hence ‘accent probabilitaire’ (‘probabilistic stress’), and depended on several factors like a word’s position and semantic weight in the utterance. He believed that Fr prominence structure was evolving from strictly oxytonic groups (final prominence) to increasingly numerous barytonic groups (initial prominence) in the media and public speeches. Vaissière (1974, 1975, cited in Vaissière 1991b) also found that speakers produced group-initial prominence in everyday speech. According to Di Cristo (1999: 163, 169), the general consensus amongst models of Fr prominence is that group-initial/‘emphatic’ and group-final prominence

are differing types of prominence (e.g. Delattre 1966a, Mertens 1991, Touati 1987), and that IPs comprise smaller rhythmic groups. Martin (1980, 1987: 934, 948) argued that Fr prosodic groups do not always correspond to (morpho-)syntactic categories, and that smaller groups with specific tonal and rhythmic structures could make an utterance sound more acceptable, even though they might divide a complete syntactic phrase. Conversely, Wenk and Wioland (1982) claimed that group-initial and group-final prominence are entirely separate phenomena, since group-initial prominence is subordinate to the basic rhythmic organisation of utterances².

2.3.2.3 Intonation

Recent descriptions of Fr intonation demonstrate the nature of f_0 movements that occur on prominent syllables, though we saw above that how exactly the Fr intonational and prominence systems are related has been controversial. (See Post (2000: 18-23) and Lacheret-Dujour and Beaugendre (1999: 100-21) for details on some early accounts of Fr intonation which identified holistic pitch contours across intonation groups rather than individual syllables (e.g. Coustenoble and Armstrong 1934, Delattre 1966a, Kenning 1979, Leon 1964)). Lacheret-Dujour and Beaugendre (1999) identified three recent approaches to modelling Fr intonation. One comprised models of how intonation links to (morpho)syntax, semantics and pragmatics, e.g. Morel's and Rossi's work (1990s), which concerned phrasal rather than syllable-level f_0 contours. Another approach concerned speech synthesis and automatic speech recognition, e.g. Vaissière's, Bailly's and Beaugendre's work (1980s-1990s). From phonetic analyses of recordings, Vaissière (1980, cited in Lacheret-Dujour and Beaugendre 1999) modelled Fr intonation as IPs comprising Prosodic words (i.e. rhythmic groups) in which final syllables had rising f_0 IP-medially, and falling f_0 IP-finally in declarative sentences; Prosodic-word-initial syllables always had a rise. In an experiment similar to the IPO perceptual approach to intonation analysis (t Hart et al. 1990), Beaugendre (1994, cited in Lacheret-Dujour and Beaugendre 1999) stylised f_0 on sixty utterances; acoustic analysis of the stylisations that listeners perceived as equivalent to the original utterances revealed that their structure was based on IP-medial (rhythmic-group-final) rises, an IP-final fall, final lengthening and pauses.

A third approach to analysing Fr intonation is the Autosegmental-Metrical framework, exemplified with the following phonological models based on phonetic data: Mertens (1987,

² Compare Fox (2000: 125): 'we might conclude that there are different kinds of accent [or prominence], each with different phonetic manifestations, and that different investigators are therefore examining different phenomena. [...] On the other hand [we might conclude] that accent does not have a consistent phonetic manifestation, and cannot, therefore, be defined in phonetic terms.'

1993)³; Jun and Fougeron (1995, 1998, 2000)⁴; Post (2000, 2002); Welby (2006). Mertens combined tonal with durational description; group-initial prominent syllables were H or L, and group-final ones were LL, LH, HH, HL, HL-, H+H+, H+L, L-L-, with double tones signifying increased duration, and H+/L- the highest/lowest extreme of the speaker's pitch range. Jun and Fougeron's model included three levels: IP, with a final boundary tone (L% or H%); Intermediate Phrase (ip), with a right-edge phrasal tone (L- or H-); Accentual Phrase (AP), transcribed as /L Hi L H*/. Generally, Hi is realised on the AP-initial prominent syllable, and H* on the AP-final prominent syllable, and f0 is interpolated between L and H targets (Fougeron and Jun 1998). However, all four tones are not always realised, depending on several factors (see Jun and Fougeron 1995), so APs can have various phonetic realisations including [HiLH*], [LHiH*] and [LHiL*] (Fougeron and Jun 1998). AP-initial prominent syllables (if prominence is realised) have rising f0, and AP-final prominent syllables can have a rising, high-level or, when IP-finally accompanied by lengthening, falling f0 (Jun and Fougeron 2000). According to Post's model, IPs can comprise the following pitch-accents (*) and boundary tones (%):

$$\left\{ \begin{array}{l} \%L \\ \%H \end{array} \right\} (H^*(L))_0 \left\{ \begin{array}{l} H^* \\ H+H^* \end{array} \right\} \left\{ \begin{array}{l} L\% \\ H\% \\ 0\% \end{array} \right\}$$

Parentheses indicate optionality, i.e. any number of IP-internal pitch-accents can occur (Post 2000: 154). The tone sequence is realised as an f0 contour temporally aligned with phonetic targets associated with certain syllables. Pitch-accented (prominent) syllables always have high f0 at some point, but can have a rise, fall, or rise-fall as f0 is interpolated between adjacent tones. For example, %L H* L H* H% has a final rise due to the preceding L and following H%, whereas %L H* H* L% has a final fall due to the preceding H* and following L% (Post 2000: 160). In Welby's analysis, which did not examine f0 falls and, unlike Post (2000), treated initial and final rises as structurally different, APs are underlyingly LHLH. The first LH is a bitonal 'edge tone' (phrase-accent/boundary-tone), whose L seeks the AP's left-hand boundary or first content word, and H is realised one or two syllables later. The second LH is a bitonal pitch-accent, with H consistently realised at the end of the AP-final syllable but L is not segmentally anchored. The first and second rises apparently correspond to group-initial and group-final prominence respectively. However, Welby (2006) argued, citing from Ladd (1996) and Vaissière (1997), that native Fr speakers may not perceive Fr AP-final rises as prominent (cf. §2.3.2.2.3), so

³ Tones were assigned to all (including non-prominent) syllables, making this analysis not quite Autosegmental-Metrical (Post 2000: 23).

⁴ Tone configurations were assigned to rhythmic units rather than individual syllables. Di Cristo and Hirst's (1984, 1986, 1993, 1997) model of Fr intonation, not detailed here, was similar in that tone configurations were assigned to whole IPs and smaller tonal units.

the Fr ‘pitch-accent’ is not completely like those in Germanic languages which associate with lexically-stressed syllables; instead both Fr rises associate with boundaries (cf. Vaissière 1991b: 117 ‘the striking fact is that in French the up and down of pitch oscillation are bounded [sic] to boundaries.’)

In sum, the final syllable of (IP-internal) rhythmic groups often has an f₀ rise, whilst IP-final syllables often have a fall or occasionally a complex contour. Rhythmic-group-initial syllables may also have a rise. Most Fr intonation accounts (though not Welby 2006) imply that these initial and final f₀ movements are associated with prominence. Intonation analyses have also found that syllable lengthening: co-occurs with f₀ movement group-finally but not group-initially (Mertens 1987, Welby 2006); is greater IP-finally than rhythmic-group-finally (Jun and Fougeron 2000); is greater in IP-finally in questions, mostly ending with a rise, than IP-finally in declarative sentences, mostly ending with a fall (Smith 2002). These findings suggest that increased duration may be a more significant prominence-lending cue than f₀ movement group-finally but not group-initially (cf. §2.3.2.2.2). Still, according to Di Cristo (1998: 217), relatively little is known about ‘tonal and temporal interplay’ in Fr.

2.3.2.4 Swiss French

The main characteristics which apparently distinguish contemporary SFr from Fr are phonological/phonetic (§2.2.2). Knecht and Rubattel (1984) and Bayard and Jolivet (1984) claimed that besides prosody, little difference exists between the two varieties. According to stereotype, SFr is slower than Fr and has unusual intonation and accentuation (Knecht and Rubattel 1984, Miller 2007). Even the following linguists made these claims, which are presumably also impressionistic, since no supporting data was offered. Grosjean et al. (2007: 2-3) stated that SFr ‘shows more pitch movement on penultimate syllables in phonological phrases than Parisian French’. Singy (2001: 271) claimed that SFr and Fr accentuation/intonation are distinct since SFr speakers tend to produce rises on penultimate syllables. In Ball’s (1997: 101) words, ‘[SFr] speakers have a tendency to accentuate the penultimate syllable of words, and to adopt a slower speed of delivery’.

The numerous views on how to best model Fr prominence and intonation demonstrate the complexity of Fr prosody, so it is unsurprising that investigation has rarely approached regional prosodic variation (cf. Miller 2007). To my knowledge, only two phonetic experiments have systematically compared SFr and Fr prosody: Miller (2007), Woehrling et al. (2008). Miller (2007) analysed speech rate and f₀ contours, the two stereotypically SFr characteristics, from recordings of six SFr speakers (Vaud canton) and six Fr speakers. For rate of read speech with pauses included, little between-variety difference occurred, but with pauses excluded, SFr speakers’ rate tended to be slower. Accentual Phrases (APs) were significantly longer in SFr than Fr. Miller (2007) concluded that longer prosodic groups separated by fewer pauses may explain

the perceived slowness of SFr. Generally, many SFr APs had a LHLH pattern with final and initial rises comparable to Fr (e.g. Jun and Fougeron 2000, Post 2000, Welby 2006). Only two SFr speakers occasionally produced fully completed rises in AP-penultimate syllables. The SFr spontaneous speech contained some instances of an IP-final rising-falling f_0 , instead of the usual fall in Fr (Miller 2007) (though Post (2000) observed rise-falls in her Fr data). Between-variety difference was observed in the alignment of tones to segments in read speech: AP-initial rises started significantly later in the first syllable in SFr than Fr, and continued into the second; AP-final rises occurred significantly earlier in SFr than Fr, so started in the penultimate and ended in the final syllable. Miller (2007: 132-33) argued that the later AP-initial rise may imply a longer time taken reaching the H target, hence perceived slower speech, and that the earlier AP-final rise could give the impression of penultimate prominence. Woehrling et al. (2008) investigated word-initial prominence and phrase-final lengthening in varieties of Fr, using a corpus of read and spontaneous speech (from the 'Phonology of Contemporary French' project, Durand et al. 2003), which included over 150 speakers from Paris, Switzerland, Alsace and Belgium. The mean segment duration was slightly higher (i.e. slower speech rate) in SFr than (Parisian) Fr, particularly in spontaneous speech. In content-word-initial syllables following a function word, neither SFr nor Fr displayed vowel lengthening, and onset consonants were longer, and substantial f_0 rises occurred more often, in SFr than Fr. In both varieties, group-final vowels were lengthened, as were SFr group-penultimate vowels, which showed more f_0 peaks than in Fr.

Some less detailed sources of data on SFr prosody exist. In a PVI experiment comparing SG and SFr monolinguals and SG~SFr bilinguals, Galloway (2007) found that the durational variability of successive intervals (both vocalic and consonantal) in SFr monolinguals' speech was similar to that found for Fr (e.g. Grabe and Low 2002). Andreassen and Detey (2007) analysed the speech produced in a sociolinguistic interview with a 31 year-old male life-long resident in the Vaud canton; they concluded that his pronunciation was generally similar to that of Fr, but phrase-penultimate syllables often had rising pitch, and the rhythm sounded relatively slow due to the realisation of phonologically long vowels. Two studies on prosody included amongst their Fr speakers one from Switzerland: Benguerel (1973) analysed Fr prominence (see §2.3.2.2.2), and Pamies Bertrán (1999) measured durational properties in a cross-linguistic rhythm study, including Fr. Presumably the Swiss participants' data were not anomalous, since neither author reported this.

Overall these findings suggest that SFr and Fr prosody are similar, but subtle differences in tonal prominence patterns and durational properties may be perceptible, contributing to a distinct Swiss accent. (Boudreault (1968) thought that the phrase-penultimate syllable lengthening he observed in Canadian Fr (as has been found in SFr) would only lead to a perceived rhythm different from Fr if this were produced systematically and considered a leftwards prominence

shift.) More research on SFr prosody production and perception is needed, to which this thesis contributes.

2.3.3 Summary (prosody)

Table 2-1 summarises the differences between SG and (S)Fr prosody identified above.

	Swiss German	French	Swiss French	
Rhythm				
In typology terms, reportedly...	'stress-timed' (though no data showing physically isochronous stress-groups)	'syllable-timed' (though data show lack of physically isochronous syllables)	no reports or data on 'syllable-timing'	
Durational variability	relatively high	relatively low	relatively low (slightly higher than Fr?)	
results from	syllable structure	complex	simple	
	phonological length contrasts	V and C	none	V only
	lexical stress	yes	no	
	reduced vowels	substantial reduction of non-prominent Vs	negligible V reduction	
Prominence				
Domain	foot/stress-group (often disyllabic)	various terms: AP/rhythmic group/prosodic word etc. (usually 2-5 syllables)		
Lexical stress	yes	no, phrasal prominence		
Location	group-initial/ left-edge/ trochaic	<i>obligatory</i> : group-final/ right-edge/iambic <i>optional (emphatic)</i> : group-initial	<i>obligatory</i> : group-final or sometimes penultimate <i>optional (emphatic)</i> : group-initial	
Phonetic correlates	dynamic f0 and increased duration, though no perceptual data	dynamic f0 and increased duration, probably lengthening primary for final prominence and f0 rise primary for initial prominence	dynamic f0 and increased duration, similar to Fr, though f0 rise and/or increased duration might be significant cue if realised on penultimate syllables (perceptual data needed)	
Table continued overleaf				

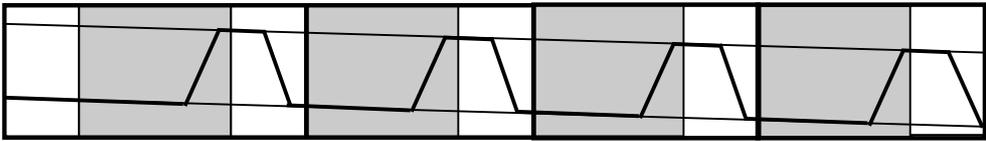
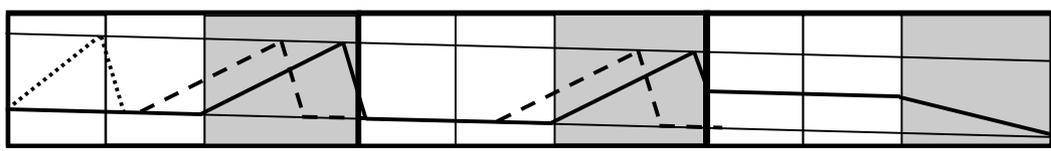
	Swiss German	French	Swiss French
Intonation			
IP-final f0 contour	often a sharp fall	often a fall	often a fall
IP-medial f0 contour	rise on prominent syllables, often starting late, continuing into next syllable	rise on (IP-non-final) prominent syllables, starting and ending in rhythmic-group-final syllable	rise on (IP-non-final) prominent syllables, starting in rhythmic-group-penultimate and continuing into final syllable
IP-initial f0 contour	rise on first content word (lexical stress)	rise sometimes on first content word, for emphasis or alternation of rhythmical pattern	
Schematic diagrams of typical IPs	 <p style="text-align: center;">Swiss German</p>		
	 <p style="text-align: center;">French ——— Swiss French - - - - (optional initial rise) ······</p> <p style="text-align: center;">shaded: prominent syllables; bold rectangles: domain of prominence; parallel lines: declination and upper bound of pitch range</p>		

Table 2-1 – Summary of SG and (S)Fr prosodic properties

Rhythm, prominence and intonation are clearly interrelated phenomena. In particular, the differences in phonological structure and phonetic manifestation of prominence between SG and Fr play a significant role in these languages' differing rhythm and intonation. The use of the term 'prominence' (not 'stress'/'accent') avoids implying that SG lexical stress, usually manifested in content-word-initial syllables, equates to Fr phrasal accent, always manifested rhythmic-group-finally. Importantly, these 'stress' and 'accent' have in common the fact that some syllables are somehow made more prominent than others. As §2.1 explained, the prosodic differences between SG and (S)Fr are an essential reason why these languages were chosen for the experiments reported in the following chapters, which investigate the perceptual interdependence of duration and f0, and add to the rarity of experimental phonetic data on SG and SFr.

The influence of dynamic f0 on the perception of duration

3.1 Summary

The experiment reported here investigates whether a dynamic f0 affects the perceived duration of non-speech sounds and isolated monosyllables, and if so, whether this depends on listeners' native language. The results demonstrate a perceived lengthening effect of dynamic f0 for all three language groups tested. This finding of a perceptual interaction between duration and f0 in a psychoacoustic task has implications for duration-based rhythm research (cf. Lehiste 1976, Rosen 1977a), and suggests that we need to investigate further these cues' interdependence in more linguistic tasks.

3.2 Previous research

Lehiste (1976) found that English speakers perceived a synthetic vowel with a rising-falling or falling-rising (i.e. dynamic) f0 as longer than one with a level f0 when both in the pair were of equal physical duration. With more English speakers, this finding was replicated using: dynamic stimuli bearing an f0 rise *or* fall (Pisoni 1976); three different vowels and non-speech sounds (Wang et al. 1976); synthetic [pa] syllables (Yu to appear).

Similar experiments did not replicate this finding. Rosen (1977b) presented Swedish listeners with synthetic vowel pairs, and found that only when the second vowel had a dynamic f0 did this perceived lengthening effect occur. No evidence for this effect was found when Rosen (1977a) played /ɛt/ stimuli with various durations and f0 contours to more Swedish listeners, who indicated whether they heard *ett* ([ɛt] 'one') or *ät* ([ɛ:t] 'eat'). Likewise van Dommelen (1991, 1993) asked German listeners to indicate which word they heard from various minimal pairs involving the /a/-/a:/ distinction (e.g. *walle-Wale* /valə/-/va:lə/, *As-Aas* /as/-/a:s/) when that vowel had a falling or level f0. For monosyllabic stimuli, falls often elicited more 'long' judgments; for disyllabic stimuli (presented in carrier phrases), falls consistently increased the number of 'short' judgments. Neither Rosen (1977a) nor van Dommelen (1991, 1993) discussed vowel quality, which may have affected their results from lexical-distinction tasks. Swedish /ɛ/ is more front and /ɛ:/ is more central (*Handbook of the IPA* 1999); German /a/ may be higher, more lax, or more front than /a:/ (see Wiese 2000: 21-22). Lehnert-LeHouillier (2007) found that Japanese but not Thai, German or Spanish speakers showed a perceived lengthening effect of falling f0. Stimuli were nonsense /tV/ monosyllables with a falling or level f0, manipulated from an Estonian speaker's production, which may be problematic if the speaker's unaspirated /t/'s did not fit within the VOT range for one phonological category expected by listeners. The stimuli were probably similar to Spanish voiceless /t/ and Thai unaspirated voiceless /t/,

possibly similar to German /d/ which is only fully voiced intervocalically, but not necessarily to Japanese /t/ which is moderately aspirated (*Handbook of the IPA* 1999).

3.3 Extending previous research

The present experiment extended previous research by addressing the methodological differences (listeners' native language, experiment design) that might have caused the conflicting findings concerning a perceived lengthening effect of dynamic f₀.

3.3.1 Listeners' native language

Lehiste (1976) suggested that in her experiment with English speakers, which appeared before those with other languages, responses possibly related to prominence perception, since f₀ movement and increased duration are both correlates of English stress. Lehnert-LeHouillier (2007) suggested that Swedes (Rosen 1977a) and Germans (van Dommelen 1993) may have behaved differently from English speakers (e.g. Lehiste 1976) because American English, unlike Swedish and German, does not have phonemic vowel-length contrasts. (Swedish and German vowel-length pairs involve quality differences, so are, arguably, similar to e.g. *bead*–*bid* /**bi:d**/–/bɪd/, though vowel-length contrasts may be less obvious in American than Southern British English.) Lehnert-LeHouillier (2007) found a perceived lengthening effect of dynamic f₀ for speakers of Japanese but not German, Thai (three languages with vowel-length contrasts) or Spanish (without vowel-length contrasts). This showed that the presence/absence of vowel-length contrasts in listeners' native language could not explain the previous studies' conflicting findings. Lehnert-LeHouillier (2007) concluded, similarly to Lehiste (1976), that Japanese speakers' responses possibly related to the interdependence of dynamic f₀ and increased duration in their perception of vowel length, since falling f₀ may occur on long, but not short, Japanese vowels.

In this thesis, a major reason for investigating SG and (S)Fr is their differing prosodic properties. Specifically in this experiment, therefore, these three language groups are well suited to testing whether native-language prosodic properties affect the way in which f₀ influences listeners' duration judgments. Furthermore, the previous similar experiments (§3.2) concerned 'standard' Germanic languages (except Lehnert-LeHouillier 2007), so speakers of non-standardised Germanic dialects and standard French varieties have not yet been tested. A summary of SG and (S)Fr prosody relevant to this experiment follows (see chapter 2 for references and details). In SG, prominent syllables have an f₀ rise, which may continue into the following perceptually non-prominent syllable, and mostly occur IP-medially, dependent on lexical stress, which mainly occurs on word-initial syllables. A late sharp f₀ fall often occurs on IP-final non-prominent syllables. Prominent syllables are lengthened, and so are IP-initial and IP-final non-prominent syllables to some extent. (S)Fr has no lexical stress; rhythmic-group-final

syllables are obligatorily prominent, generally have an f_0 fall (IP-final) or rise (IP-medial), or occasionally a rise-fall, and are extensively lengthened. Rhythmic-group-initial syllables are optionally prominent ('emphatic'), in which case they have dynamic f_0 but not increased duration. In SFr, the group-initial and group-final rises may start later and earlier respectively than in Fr.

In sum, the physical properties of lengthening and dynamic f_0 co-occur on prominent syllables in SG and (S)Fr, though not on all prominent syllables in (S)Fr and also on non-prominent syllables in SG. The question is whether these physical properties are interdependent *perceptual* cues to prominence in each language, and thus whether listeners are influenced by these properties of their native language when perceiving the length of tonally dynamic sounds, as Lehiste (1976) and Lehnert-LeHouillier (2007) suggested in terms of stress and vowel length respectively for their English and Japanese listeners. The predictions for the present experiment (§3.4) will return to this question. The above outline of differences between SG and (S)Fr should not blind us to the possibility that a perceived lengthening effect of dynamic f_0 does not depend on native language like Lehiste (1976) and Lehnert-LeHouillier (2007) interpreted from their results. We may question why listeners should refer to their native-language prosody when the task is not explicitly a linguistic one. The present experiment's results will give further insight.

3.3.2 Experiment design (previous studies)

Other factors which might explain the previous studies' conflicting findings concern their experiment design, including the task and nature of stimuli.

3.3.2.1 Listeners' task

The studies which observed this perceived lengthening effect of dynamic f_0 mostly used a forced-choice AB paradigm; listeners indicated whether stimulus A or B was longer (e.g. Lehiste 1976, Wang et al. 1976). Yu (to appear) asked listeners to rate perceived length of stimuli on a 7-point scale, where 1 was 'shortest' and 7 'longest'. For both tasks, though the stimuli were speech-like (monosyllables/vowels), the instruction required listeners to simply judge acoustically without a linguistically meaningful context. Kohler (2008: 261) suggested the term 'psychophonic' for such experiments between psychoacoustics and natural speech perception. Two studies which did not replicate the perceived lengthening effect of dynamic f_0 (Rosen 1977a, van Dommelen 1993) utilised phonemic vowel-length contrasts. Listeners indicated whether they heard word A or B, thus the instruction required them to make linguistic judgments. However, Rosen's (1977b) other task was identical to Lehiste's (1976) and did not replicate her findings. Lehnert-LeHouillier (2007) only observed this effect of dynamic f_0 in one language group, using an AXB identification procedure; listeners indicated whether stimulus X was more similar to stimulus A or B. The instruction did not explicitly require listeners to attend

to duration or to certain linguistic categories, so they may have adopted various strategies to judge similarity.

A forced-choice AB paradigm was chosen for the present experiment for several reasons. First, the results are comparable with the greatest number of previous studies, including those which did and did not find the sought effect. According to Macmillan and Creelman (1991: 134), forced-choice AB tasks are very popular because they discourage response bias and performance level is high. ABX designs have been criticised (Beddor and Gottfried 1995, Pollack and Pisoni 1971, Repp 1984), mainly because the need to compare A with X, and an intervening B with X, places a high load on memory. AXB and 4IAX tasks reduce this memory load (Repp 1984), and the experimenter does not need to explicitly instruct listeners about the dimension in which they should judge (Macmillan and Creelman 1991). However, if this lack of explicitness had led listeners to judge by non-duration cues in this experiment on duration perception, the results could have been confounded, hence a second reason for choosing an AB (over an AXB) design here. Third, a vowel-length-contrast AB design was impossible for Fr, and although SG and SFr have vowel-length contrasts, listeners would possibly attend to non-duration (e.g. spectral) cues (see comment in §3.2 on Swedish and German). Fourth, a rating-scale would have inherent problems, like some listeners never use the extreme values, and the reference points against which listeners judge the length of individual stimuli are unclear.

3.3.2.2 Linguistic versus non-linguistic stimuli

The previous studies' stimuli differed in various details. Therefore, the present experiment's stimuli were designed to investigate several variables in a single experiment, and are briefly presented now for comparison with previous studies (a full description appears in §3.5.2). Van Dommelen (1993: 383-84) suggested 'that Lehiste (1976), through the use of isolated vowels as speech material, happened to find the exception rather than the rule.' In the present experiment, two types of stimulus, linguistic and non-linguistic, were manipulated with identical f_0 contours, and presented under the same conditions to three language groups of listeners, so we can directly compare responses to both stimulus types. Previous studies involved one stimulus type and one language (e.g. Lehiste 1976, van Dommelen 1993), two stimulus types and one language (e.g. Rosen 1977a, 1977b), or one stimulus type and several languages (Lehnert-LeHouillier 2007).

Here the linguistic stimuli are [si] monosyllables. This word means 'if' or 'yes' (in response to a negative question) in Fr, and 'she' or 'they' in SG; with this frequent word, the task is arguably more linguistic than one with nonsense syllables or vowels. The phonetic properties of the voiceless fricative were particularly desirable. A plosive would entail the issue that different oral and laryngeal timing patterns occur in the realisation of (S)Fr and SG plosives (see Galloway 2007), which might affect subjects' perception of duration (a potential problem with Lehnert-

LeHouillier's (2007) study). A voiced consonant would require a decision over whether the f_0 movement should occur across the whole syllable or just the vowel. The non-linguistic stimuli are meaningless 'buzzes' with a five-formant structure. Wang et al.'s (1976) study was the only previous similar study that included non-linguistic stimuli (with a single 1500 Hz formant), the responses to which were similar to those for vowel stimuli. However, in an experiment unrelated to the perceived lengthening effect of dynamic f_0 , Lehiste (1977) found that listeners could perceive small duration differences in non-speech better than in speech stimuli, which suggests that this is worth investigating further.

3.3.2.3 Dynamic f_0 in stimuli: direction, excursion, timing

Generally, the previous studies each concentrated on one direction of f_0 movement: simple falls and/or rises, or complex contours i.e. fall-rises/rise-falls. In this experiment, stimuli include falls, rises and complex contours, and are presented under the same conditions, to test whether f_0 direction influences perceived duration. Rises and falls might have a different perceptual significance, which might depend on listeners' native language. The physical difference between a simple and a complex f_0 movement might be perceived differently in any language.

None of the previous studies varied excursion of f_0 movement, except Lehiste (1976) who found some evidence that an increase in excursion of complex f_0 contours led to a slight increase in perceived duration. The present experiment investigated this effect of complex contour excursion with other languages and different stimulus types¹.

None of the previous studies investigated timing of f_0 movement. Evidence from some pitch perception experiments suggests that this is worth pursuing. House (1990) found that timing of f_0 movement relative to spectral (in)stability affected pitch perception; listeners perceived an f_0 movement as dynamic when it occurred during a long vowel, but as a sequence of levels when it occurred during the rapid spectral changes associated with an intervocalic consonant. House (1990) argued that this was unlikely to be language-specific to his Swedish listeners. Similar findings were obtained with Fr speakers in Rossi's (1971, 1978) experiments, which investigated the threshold excursion at which an f_0 movement is perceived as dynamic rather than level. Results showed that the perceived height of a fall or rise was determined by the f_0 at a point two thirds into the vowel, i.e. once spectral stability was reached. The question of whether dynamic f_0 influences perceived duration would be less relevant in tonally dynamic stimuli that listeners perceive as non-dynamic; e.g. an f_0 fall during a consonant-vowel transition may be perceived as a low-level tone (because spectral instability reduces perceptual sensitivity to f_0 movement, and the f_0 is low once stability is reached). The present experiment includes early-

¹ The effect of excursion in simple falls and rises was tested in the pilot experiment; as there was little effect, these stimuli were later removed, and more 'filler' stimuli were then included (see §3.5.3).

and late-starting falls and rises, to test whether timing of f_0 movement relative to spectral (in)stability, which affects pitch perception, consequently influences perceived duration. Also included are stimuli with high- or low-level f_0 , to compare responses with those for stimuli with early- or late-starting f_0 movements. This further investigates House's (1990) claim that listeners recode f_0 movements in spectral instability as level tones, and tests whether pitch height influences perceived duration. Yu (to appear) found that English speakers perceived low-level stimuli ([pa] monosyllables) as shorter than mid- and high-level stimuli, and dynamic stimuli as durationally similar to high-level stimuli, when all stimuli were of equal physical duration. Yu (to appear) suggested, like Lehiste (1976), that listeners' responses possibly related to stress perception, as high f_0 and increased duration are English stress correlates.

3.3.2.4 Duration of stimuli

The previous studies included stimuli of different duration ranges. For example, Lehiste's (1976) stimuli were 270, 300 and 330 ms long, Pisoni's (1976) were 160, 200 and 240 ms long, Lehnert-LeHouillier's (2007) ranged from 222ms to 361ms in 29ms steps, and van Dommelen's (1991) ranged from 360ms to 500ms in 20ms steps. These studies drew little attention to the possibility that the physical duration of stimuli affects a perceived lengthening effect of dynamic f_0 . To investigate this further, three durations ranging across previous studies' values were assigned to the present experiment's stimuli: 250ms, 375ms, 500ms.

3.4 Hypothesis

The primary question is whether listeners with different native languages demonstrate a perceived lengthening effect of dynamic f_0 . The stimulus variables discussed above had a secondary purpose, to further explore details of previous studies, so no predictions are made concerning these. Given the previous mixed findings, predictions concerning native language are hard to make. The one previous suggestion was that a perceived lengthening effect of dynamic f_0 occurs with listeners in whose native language dynamic f_0 and increased duration are interdependent cues to prominence or vowel length (Lehiste 1976, Lehnert-LeHouillier 2007, Yu to appear). If this is correct, the following predictions could be made.

3.4.1 Swiss German

No experiments have, to my knowledge, investigated the perceptual significance of f_0 and duration in SG prominence, so we must speculate over listeners' potential behaviour. From production data, we might deduce that dynamic f_0 and increased duration are perceptually interdependent in SG, since these properties co-occur around prominent syllables. Thus we may predict that SG listeners will perceive tonally dynamic stimuli as longer than durationally equal level stimuli significantly more often than chance. However, dynamic f_0 and increased duration also co-occur around non-prominent syllables, so their interdependence as prominence cues is

unclear. It would thus be unsurprising if SG listeners did not perceive tonally dynamic stimuli as longer than durationally equal level stimuli significantly more often than chance. Moreover, studies on standard German have not found a perceived lengthening effect of dynamic f_0 (Lehnert-LeHouillier 2007, van Dommelen 1993), though SG prosody differs from standard German.

3.4.2 (Swiss) French

In perceptual experiments on Fr, Rigault (1962) found that high f_0 was a more significant prominence cue than increased duration in disyllabic words, whereas Mertens (1991) found that increased duration was a more significant prominence cue than tonal properties in short extracts of Fr. Di Cristo (1999: 162, 2000: 40) claimed that limited data (i.e. further perceptual data were needed) supported the idea that group-initial prominence is cued by tonal properties, whereas the primary cue to group-final prominence is syllable lengthening. If so, dynamic f_0 and increased duration might be independent in Fr listeners' perception, thus we may predict that Fr listeners will not perceive tonally dynamic stimuli as longer than durationally equal level stimuli significantly more often than chance. Although subtle differences between SFr and Fr prosody have been reported, no evidence suggests that these varieties differ in this characteristic that tonal properties and lengthening cue different prominence types. Thus the same prediction is made for SFr as Fr listeners.

3.4.3 Alternatives

If the suggestion is incorrect that a perceived lengthening effect of dynamic f_0 occurs with listeners in whose native language dynamic f_0 and increased duration are interdependent cues, alternative predictions are possible. All or none of these language groups could exhibit a perceived lengthening effect of dynamic f_0 ; this effect would be independent of native language, and possibly dependent on task or stimulus variables. However, the predictions in §§3.4.1–2 are retained for now, as other relevant evidence is apparently not available. The results of a pilot experiment were partially promising. Three language groups (English, German and Fr speakers) perceived tonally dynamic stimuli as longer than durationally equal level stimuli significantly more often than chance, and Fr listeners were the closest to chance.

3.5 Method

A forced-choice AB task (i.e. listeners had to indicate whether stimulus A or B was longer) was designed and run in *Praat* (Boersma and Weenik 2008-2009: version 5.0.05).

3.5.1 Subjects

All subjects (Table 3-1) reported normal hearing, and were offered a small payment. None had learned another language before obligatory foreign-language classes starting around 9-11 years (see chapter 4 for further discussion).

Language	Total	Age (years)		Sex	
		Range	Mean	Male	Female
SG	28	20–37	25.4	6	22
		18–37	24.1	7	23
SFr	30	18–33	23.8	9	19

Table 3-1 – Summary of subjects

The SG subjects were recruited in Zürich, and tested in Zürich University Phonetics Laboratory's sound-attenuated booth. They included speakers of various dialects, though most were born and brought up in Zürich so speak the Zürich dialect. All were students at Zürich University or *die Eidgenössische Technische Hochschule Zürich* (Zürich Federal Technical College). The SFr subjects were mostly recruited in Neuchâtel, and tested in a quiet room in Neuchâtel University (a few were recruited and tested in Zürich). They came from various SFr-speaking cantons, though many from Neuchâtel. Most were university students, and some were young working professionals. The Fr subjects were recruited in Cambridge, and tested in Cambridge University Phonetics Laboratory's sound-attenuated booth. They came from various regions in France. Many were resident in Cambridge, most of which were university students, on a year-long undergraduate exchange or conducting postgraduate research; others were briefly visiting Cambridge, e.g. summer-school students. None had lived in the UK before age 18.

3.5.2 Stimuli

In *Praat*, thirty-six *PitchTier* objects were created; in each a certain duration and f0 contour was specified (3 durations x 12 f0 contours: Table 3-2). Each *PitchTier* was then converted to a buzz (called 'hum' in *Praat*) using the synthesise function. This creates a pulse train, which is run through a series of filters representing five formants, and to which the f0 contour specified in the *PitchTier* is added.

	Level (L)	Falling (F) 200-100Hz	Rising (R) 100-200Hz	Complex (C)
1	100Hz 			100-200-100Hz 
2	200Hz 			100-150-100Hz 
3				200-100-200Hz 
4				200-150-200Hz 
Durations for each f0 contour shown above			250ms, 375ms, 500ms	

Table 3-2 – f0 contours and durations of stimuli

The [si] stimuli were manipulated in *Praat* from a single natural token produced by a phonetically trained male native British English speaker articulating a series of monosyllables beginning with [s] and ending in a cardinal vowel (as in e.g. Catford 1988). Therefore, no listeners heard stimuli from a speaker of their language. The realisation of a typical /i/ by SG and (S)Fr speakers may differ from the cardinal vowel [i], but not so much that these [si]'s would be lexically unrecognisable to these listeners. 100-200Hz lies within a typical male's normal pitch range (Ladefoged 2001), hence the choice of these f0 manipulations for the stimuli. The recording took place in a sound-attenuated studio using a *Marantz* PMD670 solid-state recorder and a low-noise condenser *Sennheiser* MKH40P48 microphone with a cardioid frequency response. The recording mode was set to 16 bit linear PCM, with a 44.1kHz sample rate. The file was saved as .wav format, then transferred onto a *MacBook* (Mac OS X.4) via a USB cable, and displayed in *Praat*.

One [si], 500ms long, was selected for manipulation into thirty-six different stimuli with durations and f0 contours identical to the buzzes. Naturally, f0 movement only occurred during the vowel. First the natural token was copied (copy-1). A *PitchTier* with level f0 at 100Hz was added to copy-1, which was then resynthesised (PSOLA)², resulting in a 100Hz level [si] (copy-2). Then copy-2 was copied twice (copy-3, copy-4), and a *Praat* script was run which shortened copy-3 to 75% and copy-4 to 50% during resynthesis³ (consonant and vowel by the same proportion:

² In this thesis, all resynthesis was done in *Praat*, which uses the PSOLA method (see Moulines and Charpentier 1990).

³ The following details may be of interest about how *Praat* lengthens(/shortens) sound files: It lengthens a periodic sound by copying period-sized parts of the sound, which are centred around an original glottal pulse, and windowed by a Hanning window; this window makes sure that in case of a perfectly periodic sound, the result will be identical to the original, apart from the time shift. It does not lengthen a periodic sound by copying whole glottal periods between zero crossing points, because this would give audible

Table 3-3). At that point, three [si] stimuli had been created and saved which were 500ms, 375ms and 250ms long, each with a level 100Hz f₀. This procedure was then repeated for each of the different f₀ contours (Table 3-2) by: starting with the original [si], copying it, manipulating f₀, copying this f₀-manipulated stimulus twice, and shortening both copies.

	Duration (ms)		
	Consonant	Vowel	Total
Original	206	294	500
Shortened to 75%	155	220	375
Shortened to 50%	103	147	250

Table 3-3 – Three durations of [si] monosyllables

3.5.3 Trials

Subjects heard buzz stimuli in one section of the experiment, and [si] stimuli in another, with a short break between sections; the order of sections was counterbalanced across subjects. In each trial, subjects heard two stimuli, with an inter-stimulus interval of 800 ms. Table 3-4 gives details of trials, which were presented to subjects completely randomly. Type (i) trials (shaded) relate to the hypothesis (§3.4). The order of dynamic and level stimuli within pairs was counterbalanced, since Rosen (1977b) (and, he noted, Lehiste (1976) and Pisoni (1976)) found an ordering effect. Additional trials were included for various reasons. Type (ii) trials were included to compare responses with those for early- and late-starting f₀ movements, which seek evidence for perceptual recoding of f₀ movement (§3.3.2.3). Type (iii) trials were controls with two identical level stimuli, to test whether subjects were biased in judging the first or second as longer. Type (iv) trials, about one third of the total, were ‘fillers’ comprising two stimuli with perceptibly different durations and the same level f₀, to prevent subjects becoming bored with the task and to test their accuracy in judging the longer stimulus as longer.

jitter. It lengthens an aperiodic sound by copying pieces of the original sound file with a random duration between 8 and 12 milliseconds, to prevent a ‘hum’ that would arise when lengthening a sound if all the pieces were equally long. The times of the new glottal pulses are computed first, from pitch and voicing information; for each new glottal pulse, the closest original pulse is looked up (i.e. the closest after time warping), and a window around that original pulse is copied to the new sound, centred around the time of the new pulse. (Boersma 2008)

Type	1 st stimulus	2 nd stimulus	Number of trials	
i	Are tonally dynamic stimuli judged longer than level stimuli?	L ₁	F ₍₁₋₃₎ R ₍₁₋₃₎ C ₍₁₋₄₎ 30 (10 dynamic f0 contours x 3 durations)	
		F ₍₁₋₃₎ R ₍₁₋₃₎ C ₍₁₋₄₎	L ₁ 30 (10 dynamic f0 contours x 3 durations)	
ii	Are high-level stimuli judged longer than low-level stimuli?	L ₁	L ₂ 3 (1 x 3 durations)	
		L ₂	L ₁ 3 (1 x 3 durations)	
iii	Controls	L ₁	L ₁ 3 (1 x 3 durations)	
		L ₂	L ₂ 3 (1 x 3 durations)	
iv	Fillers	L ₁ 250ms	L ₁ 375ms	36 (12 pairs x 3 occurrences)
		L ₁ 250ms	L ₁ 500ms	
		L ₁ 375ms	L ₁ 250ms	
		L ₁ 375ms	L ₁ 500ms	
		L ₁ 500ms	L ₁ 250ms	
		L ₁ 500ms	L ₁ 375ms	
		L ₂ 250ms	L ₂ 375ms	
		L ₂ 250ms	L ₂ 500ms	
		L ₂ 375ms	L ₂ 250ms	
		L ₂ 375ms	L ₂ 500ms	
		L ₂ 500ms	L ₂ 250ms	
		L ₂ 500ms	L ₂ 375ms	
Total per section (2 sections: buzzes, [si]'s)			108	
L (level), F (fall), R (rise), C (complex). Numbers 1-4 indicate specific f0 contour: see Table 3-2				

Table 3-4 – Summary of stimulus pairs (trials)

3.5.4 Procedure

Before testing began, subjects read instructions in their native language (appendix 8.1.1; German for SGs, Fr for (S)Fr). The [si] stimuli were explicitly called ‘words’, to draw subjects’ attention to the linguistic nature of these stimuli. Subjects were given chance to ask for clarification, and warned orally (and in the written instructions) that some pairs were harder than others. Subjects sat at a *MacBook* laptop (Mac OS X.4) and listened through binaural *Sennheiser* HD520 headphones. The experiment was scripted and run in *Praat*; there were two identical versions, one with German on-screen text and one with Fr on-screen text. For each trial, the following question appeared on screen: ‘Which sound was longer – sound 1 or sound 2?’. The task was to click one of two on-screen buttons labelled ‘1 was longer’ or ‘2 was longer’. There were ten practice trials per section, which included a range of stimuli (various durations, f0 contours and pair orders) from the main trials. Subjects could ask for clarification after the practice session, though none did. The total number of trials was 236 ([10 practice + 108 main] x 2 sections), and a short break occurred after every twenty. Each trial began after subjects had clicked to begin the experiment, or to register their response to the previous trial. After this click

came 800ms of silence, then the first stimulus, then 800ms of silence, then the second stimulus. In the pilot, 500ms of silence preceded each trial, but this was extended since some subjects indicated that they needed longer to focus on the next trial. Only one listening per trial was allowed, and response time was unlimited. The whole experiment lasted approximately twenty-five minutes.

3.5.5 Analysis

Responses were recorded in *Praat* and transferred to *Excel*, in which counts of each subject's responses to the various trial types (Table 3-4) were made. Analysis consisted of three stages which addressed:

- 1) responses to 'filler' and control stimuli.
- 2) the main question related to the hypothesis: do subjects perceive tonally dynamic stimuli as longer than level stimuli significantly more often than chance, and do language groups differ?
- 3) other variables in the design.

The software used was *Excel* for binomial tests, *SPSS* for ANOVAs and the *R* software environment (<http://www.r-project.org/>) for regression analyses. Before each statistical analysis, all data were explored graphically, and (where appropriate) tests of normality (Shapiro-Wilk, skewness and kurtosis statistics) and homogeneity of variance (Levene) were conducted. Unless otherwise stated, the data did not violate the assumptions of normality or homogeneity of variance.

3.6 Results

3.6.1 Level stimuli: fillers and controls

Fillers had a 'correct' response, since one stimulus was physically longer than the other. According to binomial probability, a subject needed to have at least 69.4% correct (i.e. a maximum of 11 errors in 36 filler stimuli) to score significantly above the level of chance (probability=0.5, $p=0.01$). All individuals scored highly significantly above chance (see Table 3-5), so none was rejected for failing to complete the task accurately.

Language	Buzzes		[si] stimuli	
SG ($k=28$)	\bar{x}	95.34	\bar{x}	96.43
	s	4.21	s	4.53
SFr ($k=30$)	\bar{x}	95.93	\bar{x}	96.57
	s	4.11	s	4.04
Fr ($k=28$)	\bar{x}	93.45	\bar{x}	93.85
	s	4.42	s	5.20

Table 3-5 – Mean and standard deviation (across k subjects) of correct ‘filler’ responses

The control trials were pairs of identical level stimuli, testing whether there was a bias to perceive the first or second as longer. Figure 3-1 shows that SGs responded ‘1st sound is longer than 2nd sound’ (1st>2nd) above chance (50%) for both types of sound, whereas (S)FrS responded around chance for [si] stimuli but similarly to the SGs for buzzes. To test for significance, binomial probabilities were approximated from the standard normal distribution. Table 3-6 shows that SG responses were significantly above chance, whereas Fr responses were not, and SFr responses were significantly above chance for buzzes but not [si] stimuli.

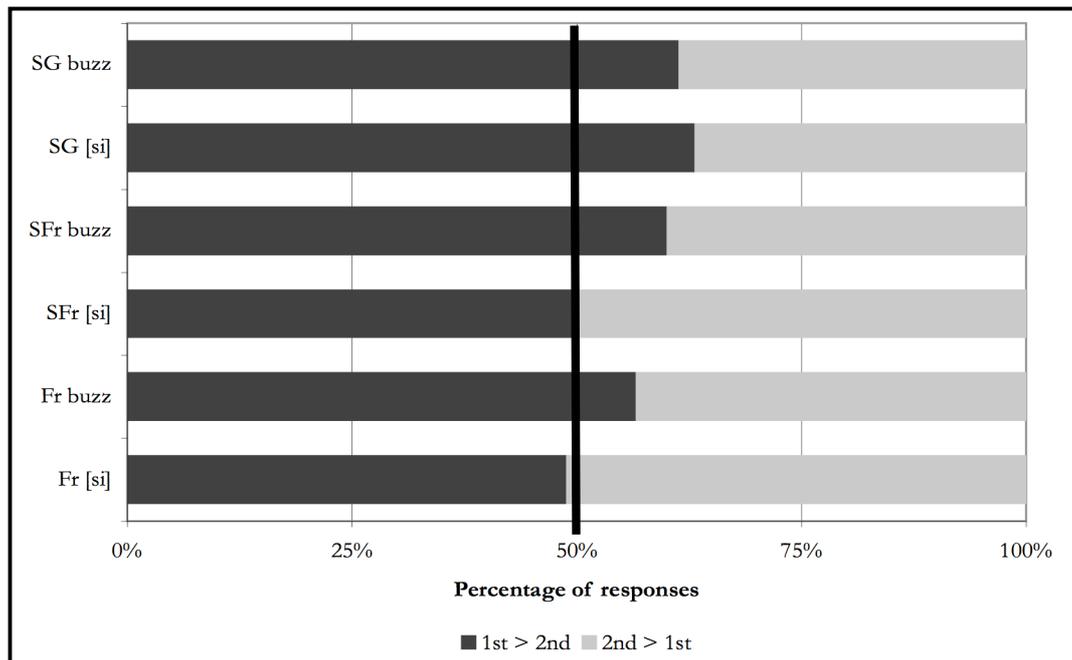


Figure 3-1 – Percentage of ‘1st sound is longer than 2nd sound’ responses compared to the level of chance

Language	Buzzes				[si] stimuli			
SG ($k=28$)	n	103	N	168	n	106	N	168
	z	2.93	p	<0.01*	z	3.40	p	<0.001*
SFr ($k=30$)	n	108	N	180	n	90	N	180
	z	2.68	p	<0.01*	z	0.00	p	>0.05 (non-sig.)
Fr ($k=28$)	n	95	N	168	n	82	N	168
	z	1.70	p	>0.05 (non-sig.)	z	0.31	p	>0.05 (non-sig.)

k , number of subjects; n , number of '1st>2nd' responses; N , total number of trials; z , z-score for the binomial probability approximated from the standard normal distribution; * significant

Table 3-6 – '1st sound is longer than 2nd sound' responses, and significance tests

These results suggest that native language and type of sound may affect perceived duration. These factors were investigated further by analysing responses to trials which had one dynamic and one level stimulus, the order of which was counterbalanced, so the bias (in some subjects) for choosing first stimuli is not problematic.

3.6.2 Dynamic stimuli

Figure 3-2 displays data concerning the main question: did each language group respond 'dynamic stimulus is longer than level stimulus' (henceforth 'D>L') significantly more often than chance? According to binomial probabilities approximated from the standard normal distribution, in all language groups and for both types of sound, the number of 'D>L' responses was significantly above chance (50%) (Table 3-7).

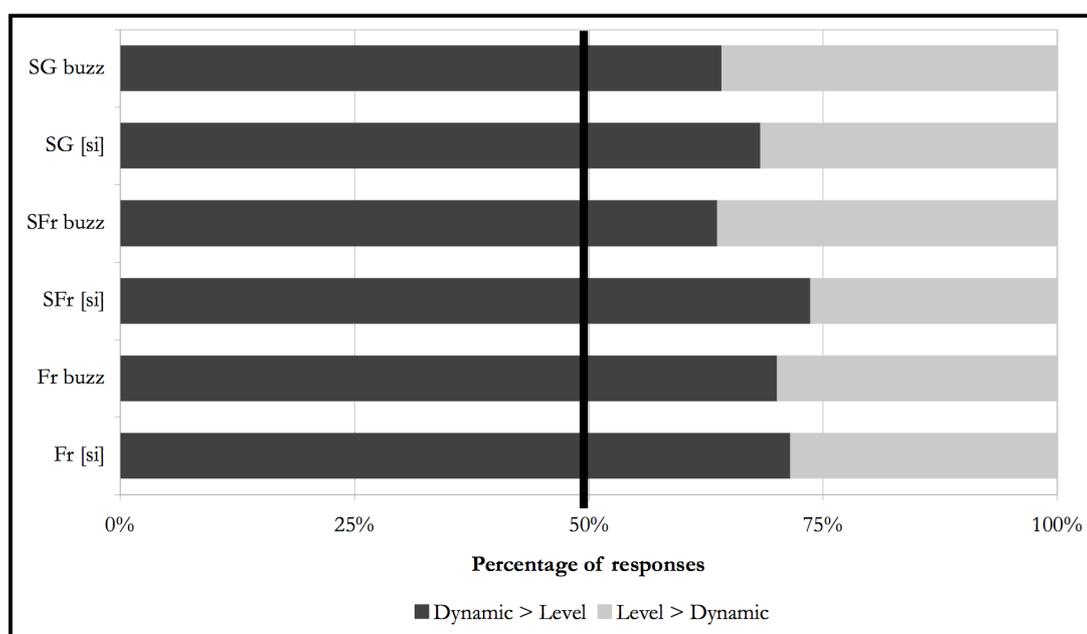


Figure 3-2 – Percentage of 'dynamic is longer than level' ('D>L') responses compared to the level of chance

Language	Buzzes				[si] stimuli			
SG ($k=28$)	n	1078	N	1680	n	1147	N	1680
	z	11.61	p	<0.0001	z	14.98	p	<0.0001
	\bar{x}	64.17	s	15.72	\bar{x}	68.27	s	12.91
SFr ($k=30$)	n	1146	N	1800	n	1325	N	1800
	z	11.60	p	<0.0001	z	20.03	p	<0.0001
	\bar{x}	63.67	s	16.50	\bar{x}	73.61	s	10.17
Fr ($k=28$)	n	1177	N	1680	n	1201	N	1680
	z	16.44	p	<0.0001	z	17.61	p	<0.0001
	\bar{x}	70.06	s	19.83	\bar{x}	71.49	s	15.22
n , number of ‘D>L’ responses; N , total number of trials; z , z-score for the binomial probability approximated from the standard normal distribution; \bar{x} (s), mean (standard deviation) percentage of ‘D>L’ responses across k subjects								

Table 3-7 – ‘Dynamic is longer than level’ (‘D>L’) responses, and significance tests

A two-way mixed-measures ANOVA with the factors *sound* (buzz, [si]: repeated-measures) and *language* (SG, SFr, Fr: between-groups) was conducted on the percentage of ‘D>L’ responses. There was a main effect for *sound* [$F(1,83)=9.075, p=0.003$], but no interaction of *sound* \times *language* [$F(2,83)=2.193, p>0.05$], nor a main effect for *language* [$F(2,83)=0.848, p>0.05$]. For the control stimuli, *sound* and *language* both had an effect (§3.6.1), whereas in this initial analysis for dynamic stimuli, there was a significant difference in responses to buzzes and [si] stimuli, but not between language groups.

3.6.3 Further analysis

Several variables were included in the experiment design for comparison with previous studies. We have already seen that the type of sound (buzzes or [si] monosyllables) had a significant effect. To explore whether other variables had an effect, a logistic regression analysis was conducted on the entire data-set. Since each subject responded to several stimuli, a mixed model with random and fixed effects must be fitted (Baayen 2008, Garson 2009). (For explanation of mixed models, see Faraway 2006, Baayen 2008, Baayen et al. 2008). The random effect was *subject*, which introduced adjustments to the intercept grouped by each subject (Baayen et al. 2008: 8). The fixed effects were *native language* (3 levels), *test order* (2 levels), *sound* (2 levels), *f0 direction* (3 levels), *f0 timing* (3 levels), *duration* (3 levels), *dynamic order* (2 levels). Test order refers to whether the buzzes or [si] monosyllables were heard first; dynamic order refers to whether the dynamic stimulus came first or second in each trial. The dependent variable was *response* (‘D>L’ =

1; ‘L>D’ = 0). Few statistical packages offer generalised linear mixed modelling. The R software environment was used, since the required model could be specified within the ‘lmer’ function⁴.

Table 3-8 displays the output; the right-most column, which displays significance values for fixed effects, is of most interest. We see that five fixed effects were (almost) significant: sound, direction of f0 movement, timing of f0 movement, duration, and order of dynamic sound in pair. Two fixed effects were non-significant: native language and test order.

	<i>Fixed effects</i>	Estimate	Std. Error	χ	p
	Intercept	0.460	0.167	2.756	0.006**
native language	SG				
	SFr	0.094	0.184	0.513	0.608
	Fr	0.302	0.188	1.608	0.108
test order	buzzes 1 st				
	[si]’s 1 st	0.060	0.151	0.394	0.694
sound	buzz				
	[si]	0.282	0.044	6.353	<0.0001***
f0 direction	fall				
	rise	-0.138	0.057	-2.419	0.016*
	complex	0.017	0.067	0.258	0.796
f0 timing	total				
	late	0.124	0.070	1.768	0.077 ^m
	early	0.036	0.070	0.523	0.601
dynamic order	dynamic 1 st				
	dynamic 2 nd	-0.300	0.044	-6.750	<0.0001***
duration	250ms				
	375ms	0.469	0.054	8.657	<0.0001***
	500ms	0.376	0.054	7.008	<0.0001***
<ul style="list-style-type: none"> • p-values were calculated using the ‘Markov Chain Monte Carlo’ method. Hornik (2008) argues that this is the best option for this model and sample number. In this thesis, all regression models run in R used this method. • significance: *** $p < 0.0001$; ** $p < 0.01$; * $p < 0.05$; ^m marginally significant, $p < 0.1$ 					

Table 3-8 – Output of regression model (‘D>L’ responses)

⁴ The formula was:

```
> expt1dynamic.lmer = lmer(response ~ language + test_order + sound + direction + timing +
duration + dynamic_order + (1 | subject), data = expt1dynamic, family = "binomial")
```

For space reasons, the effects of shape and excursion of complex contours were not examined further. A graphical inspection of the responses to complex-contour stimuli demonstrated no effect of excursion, and little difference between rise-falls and fall-rises for [si] stimuli, though fall-rises showed a slightly greater perceived lengthening effect than rise-falls for buzzes. The following sections report further analyses of the five variables with significant effects (from Table 3-8), which are interesting, but should not distract from the overall finding of a language-independent perceived lengthening effect of dynamic f0.

3.6.3.1 Direction of f0 movement

The regression analysis (Table 3-8) revealed that rising f0 was a significantly worse predictor of a 'D>L' response than falling f0, and that complex contours and falls were not significantly different as predictors. Figure 3-3 shows the mean and standard deviation of the percentages of 'falling/rising/complex is longer than level' responses ('F>L'; 'R>L'; 'C>L').

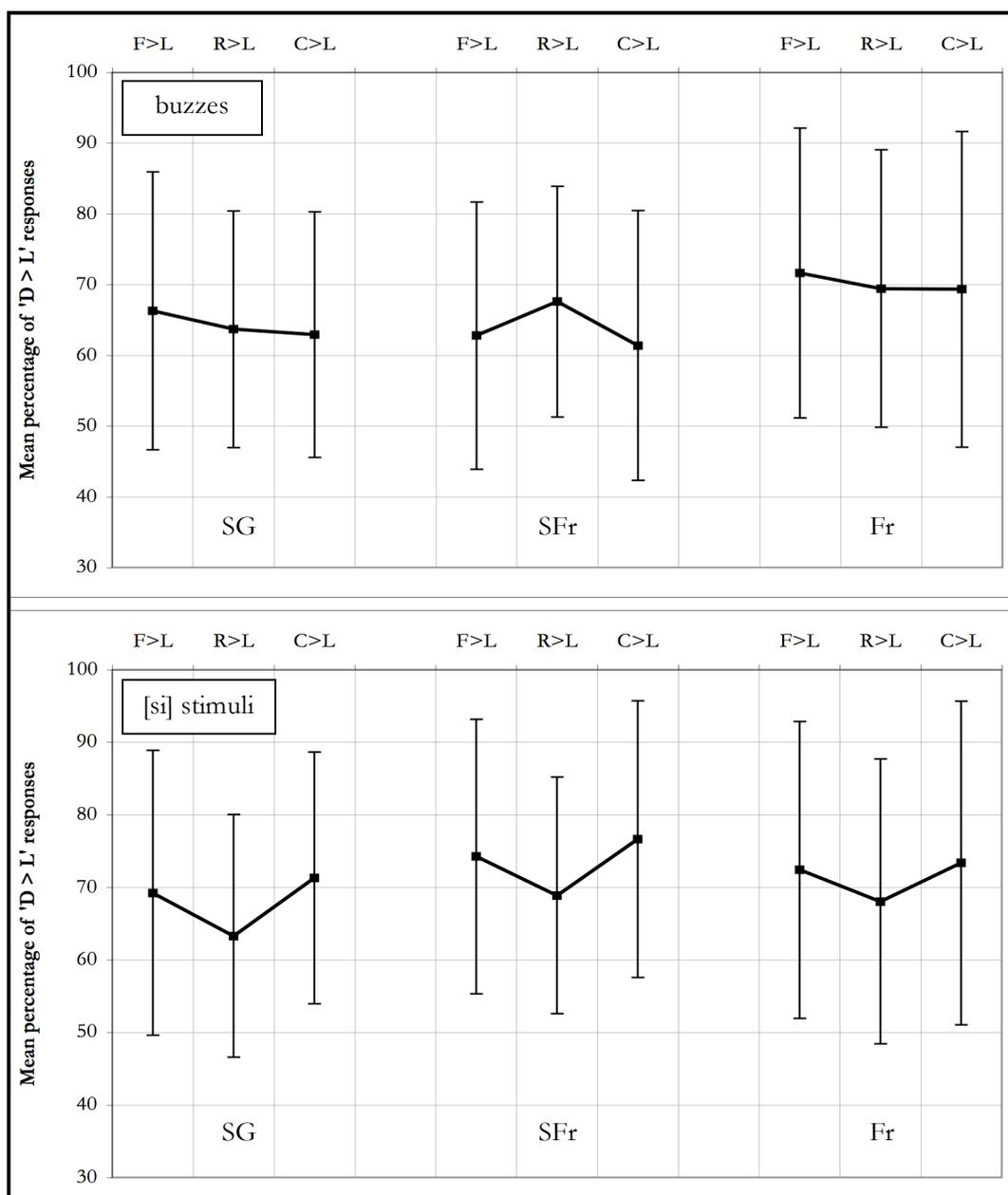


Figure 3-3 – Percentage (out of total number possible) of ‘D>L’ responses depending on direction of f_0 change; mean across 28 SG, 30 SFr, 28 Fr subjects (error bars: ± 1 standard deviation)

The pattern of results is different for buzzes and [si] stimuli, but similar across languages. A three-way mixed-measures ANOVA was calculated with the factors *sound* (buzz, [si]), *direction* (fall, rise, complex) (both repeated-measures) and *language* (SG, SFr, Fr: between-groups). There were main effects of *direction* and *sound*, but not *language* (Table 3-9).

Source	df	Mean square	F	p
sound	1	0.286	7.692	0.007**
sound × language	2	0.078	2.090	0.130
Error	83	0.037		
direction	1.844	0.038	3.416	0.039*
direction × language	3.687	0.007	0.644	0.619
Error	153.017	0.011		
sound × direction	1.913	0.099	12.497	<0.0001***
sound × direction × language	3.825	0.010	1.237	0.298
Error	158.751	0.008		
language	2	0.015	0.874	0.421
Error	83	0.017		
<ul style="list-style-type: none"> • A Greenhouse-Geisser correction was applied since sphericity could not be assumed (Mauchly's test, $p < 0.05$) • significance: *** $p < 0.0001$; ** $p < 0.01$; * $p < 0.05$ 				

Table 3-9 – ANOVA output: *sound × direction × language*

Within-groups contrasts for the significant interaction revealed that it resulted from the difference between ‘F>L’ and ‘R>L’ responses [$F(1,83)=7.558$, $p=0.007$], and the difference between ‘R>L’ and ‘C>L’ responses [$F(1,83)=31.025$, $p<0.0001$] in the two *sound* conditions. Listeners perceived falls and complex f0 contours as longer than level significantly more often when hearing [si] stimuli rather than buzzes; type of sound hardly affected responses to rises, which were perceived as longer than level significantly less often than falls and complex contours were.

3.6.3.2 Timing of f0 movement

Falls and rises included three different timings of f0 movement: ‘total’ (occurred across entire buzz/vowel); late (started at the buzz/vowel mid-point); early (ended at the buzz/vowel mid-point). Complex contours always occurred across the entire buzz/vowel. The regression analysis (Table 3-8) revealed that late f0 movements were an almost significantly better predictor of a ‘D>L’ response than ‘total’ movements. A visual inspection of the data suggested that the interaction of f0 timing and sound (buzz or [si]) was worth investigating further. Figure 3-4 displays the mean and standard deviation of the percentages of ‘late fall/early fall/late rise/early rise is longer than level’ responses (‘LF>L’; ‘EF>L’; ‘LR>L’; ‘ER>L’). These data were not split

into language groups, because another variable (language) would complicate the interpretation of an already potentially complex interaction.

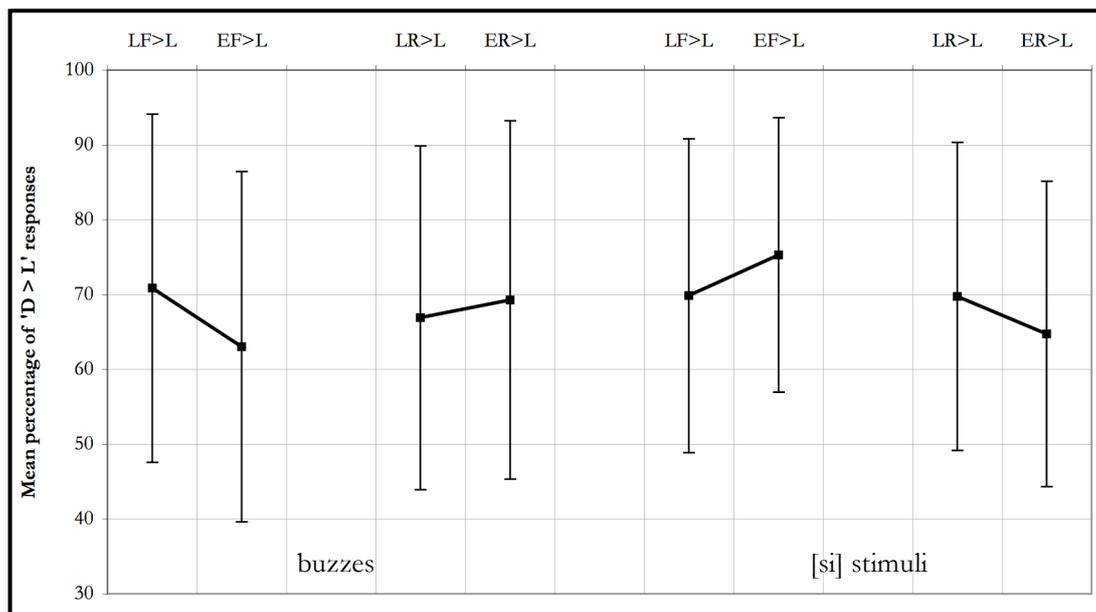


Figure 3-4 – Percentage (out of total number possible) of ‘D>L’ responses depending on timing and direction of f0 change; mean across all subjects (error bars: ± 1 standard deviation)

A three-way repeated-measures ANOVA was calculated with the factors *direction* (fall, rise), *timing* (late, early) and *sound* (buzz, [si]). According to Shapiro-Wilk tests, this subset of the data was not as normally distributed as the data used in the previous ANOVAs. However, skewness and kurtosis statistics when divided by the standard error resulted in values between ± 1.96 for most data-series in the subset, indicating a non-significant deviation from the normal distribution ($p > 0.05$) (Field 2005). Even with the few deviant data-series, ANOVA is robust against violation of the assumption of normally distributed data (Field 2005). There were no main effects of *direction* [$F(1,85)=2.437, p > 0.05$], *timing* [$F(1,85)=0.759, p > 0.05$] or *sound* [$F(1,85)=1.629, p > 0.05$], but there were two significant interactions: *direction* \times *sound* [$F(1,85)=7.393, p=0.008$] and *direction* \times *timing* \times *sound* [$F(1,85)=14.899, p < 0.0001$]. The two-way interaction corresponds to the previous ANOVA with a *direction* \times *sound* interaction (Table 3-9). Listeners perceived falls as longer than level significantly more often when listening to [si] stimuli rather than buzzes, but type of sound hardly affected responses to rises. The three-way interaction reveals a more complex situation when we add the factor of f0 timing. For buzzes, listeners perceived late falls as longer than level significantly more often than early falls, but late rises as longer than level significantly less often than early rises. Conversely for [si] stimuli, listeners perceived late falls as longer than level significantly less often than early falls, but late rises as longer than level significantly more often than early rises. Thus, as Figure 3-4 illustrates, how often a stimulus was

perceived as longer than its level counterpart is a mirror image in each *sound* condition and each *direction* condition, when the timing of f0 movement is considered.

3.6.3.3 Duration of stimuli

For each of the ten different dynamic f0 contours, there were three different durations of stimulus, each paired with a level stimulus of equal duration (250ms, 375ms, 500ms). The regression analysis (Table 3-8) revealed that durations of 375ms and 500ms were significantly better predictors of a ‘D>L’ response than a duration of 250ms. Figure 3-5 shows the mean and standard deviation of the percentages of ‘D>L at 250ms/375ms/500ms’ responses (‘250D>L’; ‘375D>L’; ‘500D>L’).

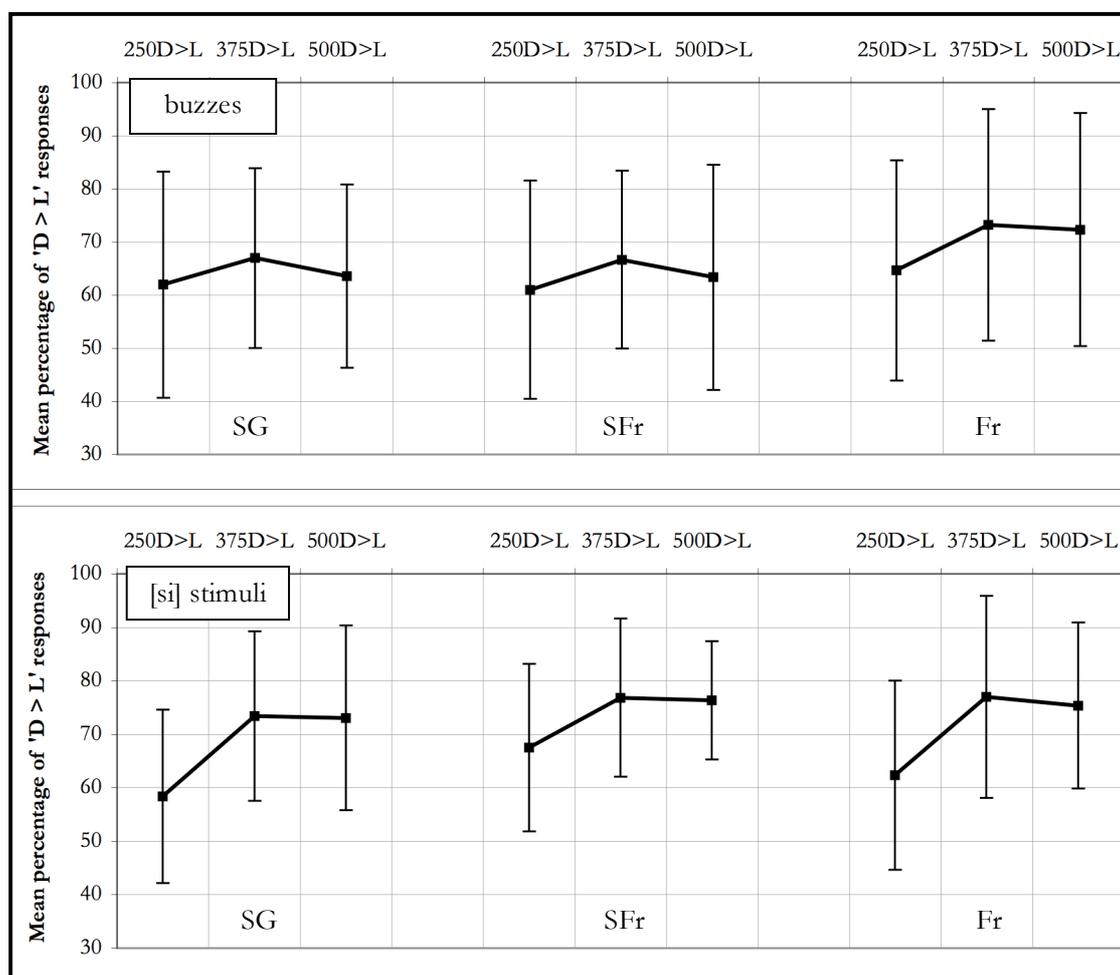


Figure 3-5 – Percentage (out of total number possible) of ‘D>L’ responses depending on duration of stimuli; mean across 28 SG, 30 SFr, 28 Fr subjects (error bars: ± 1 standard deviation)

The pattern of results is similar for buzzes and [si] stimuli, and for all language groups. A three-way mixed-measures ANOVA was calculated with the factors *sound* (buzz, [si]), *duration* (250ms, 375ms, 500ms) (both repeated-measures) and *language* (SG, SFr, Fr: between-groups). There were main effects of *sound* and *duration*, but not *language* (Table 3-10).

Source	df	Mean square	F	p
sound	1	0.343	8.996	0.004**
sound × language	2	0.081	2.119	0.127
Error	83	0.038		
duration	1.841	0.502	28.365	<0.0001***
duration × language	3.682	0.011	0.602	0.648
Error	152.803	0.018		
sound × duration	1.735	0.095	7.405	0.002**
sound × duration × language	3.470	0.009	0.677	0.588
Error	144.017	0.013		
language	2	0.015	0.860	0.427
Error	83	0.017		
<ul style="list-style-type: none"> • A Greenhouse-Geisser correction was applied since sphericity could not be assumed (Mauchly's test, $p < 0.05$) • significance: *** $p < 0.0001$; ** $p < 0.01$ 				

Table 3-10 – ANOVA output: *sound* × *duration* × *language*

Within-groups contrasts for the significant interaction revealed that it resulted from the difference between ‘250D>L’ and ‘375D>L’ responses [$F(1,83)=7.209$, $p=0.009$], and the difference between ‘250D>L’ and ‘500D>L’ responses [$F(1,83)=10.809$, $p=0.001$] in the two *sound* conditions. Listeners perceived 375ms and 500ms dynamic stimuli as longer than level significantly more often when hearing [si] stimuli rather than buzzes, whereas type of sound hardly affected responses to 250ms dynamic stimuli, which were perceived as longer than level significantly less often than 375ms and 500ms dynamic stimuli were. Note that the pitch-carrying part of each [si] stimulus was shorter than in the equivalent buzz.

3.6.3.4 Order of stimuli

There were two points concerning ordering in the experiment: ‘test order’ – half the subjects completed the buzz section first, and half the [si] section first; ‘dynamic order’ – for each pair of one dynamic followed by one level stimulus there was a counterpart with one level followed by one dynamic stimulus. The regression analysis (Table 3-8) revealed that dynamic order (but not test order) was a significant predictor of a ‘D>L’ response. This effect of order within each pair does not affect the experiment’s findings, because it was cancelled out by a counterbalanced design, but as it is an interesting finding, it was explored further. Figure 3-6

shows the mean and standard deviation of the percentages of 'D>L when dynamic is first/second' responses ('D1>L'; 'D2>L').

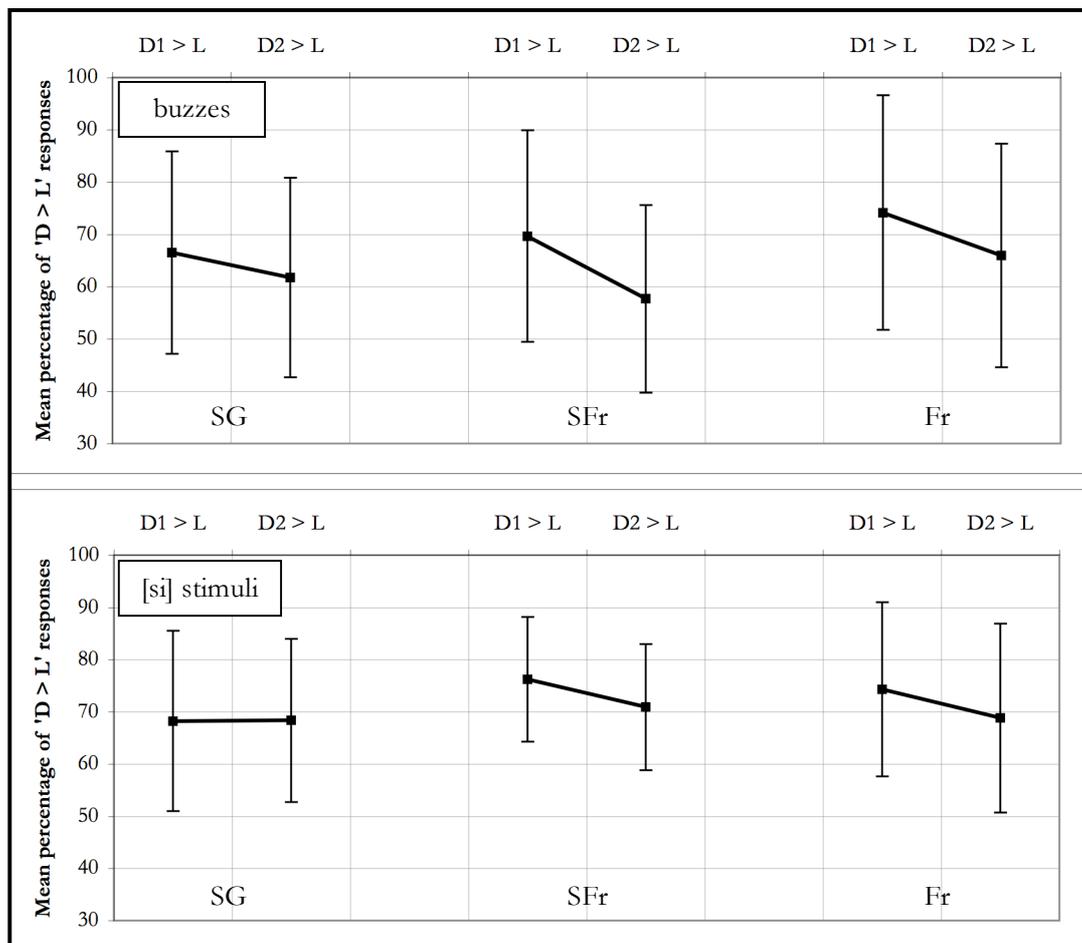


Figure 3-6 – Percentage (out of total number possible) of 'D>L' responses depending on whether the dynamic stimulus was heard first or second; mean across 28 SG, 30 SFr, 28 Fr subjects (error bars: ± 1 standard deviation)

A three-way mixed-measures ANOVA was calculated with the factors *sound* (buzz, [si]), *order* (dynamic 1st, dynamic 2nd) (both repeated-measures) and *language* (SG, SFr, Fr: between-groups). There were main effects of *sound* and *order*, but not *language*, and a significant interaction for *sound* \times *order* (Table 3-11). A dynamic stimulus was judged longer than its level counterpart significantly more often if the dynamic stimulus came first rather than second. If the sounds were [si]'s, the difference between the two orders was much lower than if the sounds were buzzes.

Source	df	Mean square	F	p
sound	1	0.229	8.996	0.004**
sound × language	2	0.054	2.119	0.127
Error	83	0.025		
order	1	0.304	13.144	<0.0001***
order × language	2	0.031	1.325	0.271
Error	83	0.023		
sound × order	1	0.049	4.369	0.040*
sound × order × language	2	0.003	0.251	0.779
Error	83	0.011		
language	2	0.015	0.860	0.427
Error	83	0.017		
significance: *** $p < 0.0001$; ** $p < 0.01$; * $p < 0.05$				

Table 3-11 – ANOVA output: *sound × order × language*

It is unsurprising to find such an ordering effect given that some subjects' responses to the control stimuli revealed a bias towards perceiving the first stimulus as longer (§3.6.1). In the controls, the first-stimulus bias was lower for [si] stimuli than buzzes; similarly, in the dynamic [si] stimuli, there was little difference in the number of 'D>L' responses between dynamic-stimulus-first and dynamic-stimulus-second trials. No significant interaction of order and language occurred for responses to dynamic stimuli, unlike for the control stimuli. In fact the SGs, who were most biased towards perceiving the first sound as longer in the controls, showed the least difference in number of 'D>L' responses between the two dynamic order conditions.

3.6.3.5 f0 height

Several trials were included which had one low-level (100Hz) and one high-level (200Hz) stimulus. Figure 3-7 shows that the percentage of 'high is longer than low' (henceforth 'high>low') responses differed for the [si] and buzz stimuli. According to binomial probabilities approximated from the standard normal distribution, subjects in all language groups responded 'high>low' at a level significantly above chance (50%) for buzzes but not [si] stimuli (Table 3-12).

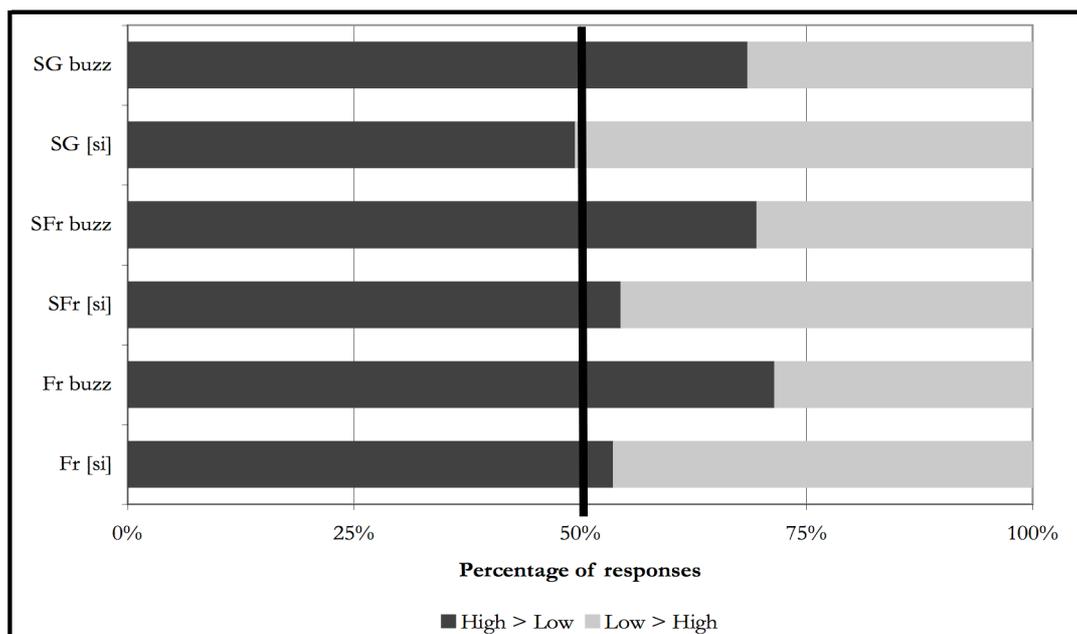


Figure 3-7 – Percentage of ‘high>low’ responses compared to the level of chance

Language	Buzzes				[si] stimuli			
SG	<i>n</i>	115	<i>N</i>	168	<i>n</i>	83	<i>N</i>	168
(<i>k</i> =28)	\bar{z}	4.78	<i>p</i>	<0.001*	\bar{z}	-0.15	<i>p</i>	>0.05 (non-sig.)
SFr	<i>n</i>	125	<i>N</i>	180	<i>n</i>	98	<i>N</i>	180
(<i>k</i> =30)	\bar{z}	5.22	<i>p</i>	<0.001*	\bar{z}	1.19	<i>p</i>	>0.05 (non-sig.)
Fr	<i>n</i>	120	<i>N</i>	168	<i>n</i>	90	<i>N</i>	168
(<i>k</i> =28)	\bar{z}	5.56	<i>p</i>	<0.001*	\bar{z}	0.93	<i>p</i>	>0.05 (non-sig.)
<i>k</i> , number of subjects per group; <i>n</i> , number of ‘high>low’ responses; <i>N</i> , total number of trials; \bar{z} , z-score for the binomial probability approximated from the standard normal distribution; * significant								

Table 3-12 – ‘High>low’ responses, and significance tests

For each sound type, three pairs of a high-level followed by a low-level stimulus were included, one pair at 250ms, 375ms and 500ms, and each pair had a counterpart with the low-level followed by the high-level stimulus. To explore the effect of the variables *duration* and *order*, a logistic regression analysis was conducted on this subset of responses to high and low stimuli. A mixed model with random and fixed effects was fitted using the R software environment (as in §3.6.3). The fixed effects were *native language* (3 levels), *sound* (2 levels), *order* (2 levels), *duration* (3 levels). The random effect was *subject*, which introduced adjustments to the intercept grouped by

each subject. The dependent variable was *response* ('high>low' = 1, 'low>high' = 0)⁵. The output accords with that for the dynamic stimuli: type of sound, order and duration, but not native language, were significant predictors of a 'high>low' response (right-most column of Table 3-13).

	<i>Fixed effects</i>	Estimate	Std. Error	χ	<i>p</i>
	Intercept	0.322	0.184	1.752	0.078
native language	SG				
	SFr	0.173	0.184	0.936	0.349
	Fr	0.091	0.184	0.496	0.620
sound	buzz				
	[si]	-0.676	0.131	-5.154	<0.0001***
order	low 1 st				
	high 1 st	0.423	0.131	3.235	0.001**
duration	250ms				
	375ms	0.101	0.157	0.645	0.519
	500ms	0.525	0.162	3.248	0.001**
significance: *** $p < 0.0001$; ** $p < 0.01$					

Table 3-13 – Output of regression model ('high>low' responses)

3.7 Discussion

The hypothesis (§3.4) was clearly tentative. It was predicted that SG listeners might perceive tonally dynamic stimuli as longer than durationally equal level stimuli significantly more often than chance (though perhaps not), whereas (S)Fr listeners would not perceive tonally dynamic stimuli as longer than durationally equal level stimuli significantly more often than chance. The results provide evidence supporting the first prediction for SG, but against the prediction for (S)Fr. All language groups perceived dynamic stimuli as longer than level stimuli significantly more often than chance; in none of the statistical analyses did a cross-linguistic difference approach significance. Two aspects of these results must be separated. The first involves individual languages as linguistic systems (i.e. SG, SFr, Fr) and the behaviour of their speakers; the second concerns human language in general. These are now discussed in turn.

⁵ The formula was:

```
> expt1level.lmer = lmer(response ~ language + sound + order + duration + (1|subject), data =
expt1level, family = "binomial")
```

3.7.1 Individual languages

Before this experiment, we could reason that listeners' native language might explain the similar previous studies' sometimes conflicting results. A reason for investigating SG and (S)Fr was their prosodic differences: if a perceived lengthening effect of dynamic f0 depends on the prosodic characteristics of listeners' native language, it would be likely to manifest itself in these data. However, such an effect was found to be independent of native language, since all groups responded similarly. Listeners' response behaviour apparently did not reflect their native-language prosody. Consider the data for the effect of f0 direction, along with SG and (S)Fr prosodic properties. Rises were perceived as longer than level less often than falls and complex contours were. In both languages, substantial f0 falls mainly occur IP-finally, where there is also evidence for phrase-final lengthening. Perhaps for this linguistic reason listeners perceptually associated falling f0 with increased duration. Yet if this perceptual association of increased duration and f0 movement applied ubiquitously, we could expect rises to have a similarly large effect, since (like falls) rising f0 and increased duration co-occur in both languages. In SG, prominent syllables have rising f0 and are typically lengthened; in Fr, IP-medial prominent syllables often have rising f0 and are lengthened. However, this experiment's perceptual data show that rises clearly have less effect than falls, though the perceived lengthening effect of rising f0 is still well above chance.

This asymmetry between rises and falls has been observed by others. Ohala (1978) reported evidence that falls needed a greater excursion than rises to be perceived as equally prominent, though Rossi (1971, 1978) presented data that falls and rises were equally perceptible. Yet, if anything, the present experiment's data suggest the opposite, that falls might be more perceptually prominent (thus judged longer than level more often) than equal-sized rises. Ohala (1978) also reported asymmetry in production, claiming that 'speakers are able to produce a falling pitch over a given pitch interval much faster than a rising pitch over the same interval.' Perhaps the present experiment's listeners perceived the rising stimuli as fast (compared to the falling ones) for f0 to increase that much in that time, and this concept of speed translated to 'short duration' in their responses. Although rises and falls physically co-occur with lengthening in SG and (S)Fr, how dynamic f0 and increased duration interact in perception is clearly complicated.

The idea that listeners respond according to their native language's prosodic characteristics was proposed by previous authors. Lehiste (1976) suggested that her English speakers may have perceived dynamic stimuli as longer than level ones because dynamic f0 and increased duration are English stress correlates. Lehnert-LeHouillier (2007) suggested that her Japanese speakers may have perceived stimuli with falls as longer than level ones because falling f0 is a correlate of long, but not short, Japanese vowels. We may question why listeners should relate their responses to language-specific prosody in a psychoacoustic task with non-linguistic

(buzz) stimuli. However, some evidence for this exists. Iversen et al. (2009) found that native-language properties may have affected English and Japanese listeners' grouping of non-linguistic pure-tone stimuli; but Hay and Diehl (2007) found no evidence for native-language influence in a similar experiment with French and English listeners (see chapter 1). From brain-imaging, Salmelin et al. (1999) found evidence for influence of native-language segmental properties on non-linguistic stimulus processing: Germans showed stronger responses to pure tones than Finns in terms of magnetoencephalography (MEG) activations in the auditory cortices. Salmelin et al. (1999) suggested that this difference might arise from the fact that /i/, /a/ and /u/ cover a smaller formant-frequency range in German than in Finnish, so the frequency resolution for language processing has developed to be higher in Germans' than Finns' brains, hence a greater MEG activation trace. Even if a cross-linguistic difference in behavioural responses is not evident in the present experiment's data, we do not know whether there were subtle differences in how the listeners cognitively processed the stimuli.

It may be more surprising that listeners did not relate their responses to language-specific prosody in the 'psychophonetic' (Kohler 2008) task with linguistic [si] stimuli, which the instructions explicitly referred to as 'words'. However, listeners heard these stimuli isolated from linguistic context. Furthermore, the 500ms (but not the 250ms or 375ms) stimuli were durationally equal to the original citation-form syllable, thus unusually long, and the f0 contours were more stylised than we find naturally in these languages. This was necessary for a controlled cross-linguistic experiment, since it is unclear which specific shapes of f0 contour are perceptually equivalent in each language. With repeated exposure to many [si] pairs, listeners may not have treated the task as more specific to their language than the task with buzzes. Yet listeners responded differently overall to buzzes and [si] stimuli, even if not in a way which reflected listeners' native language. This links to the second aspect of the results concerning human language in general.

3.7.2 Human language

A highly significant difference in responses was observed between the buzzes and [si] stimuli in all statistical analyses (except the *sound* × *f0 direction* × *timing* ANOVA). Listeners more often perceived a dynamic sound as longer than level when listening to linguistic [si] stimuli rather than non-linguistic buzzes. Furthermore, type of sound showed significant interactions with f0 direction, stimulus duration, and order of dynamic/level stimuli. Listeners perceived falls and complex f0 contours as longer than level more often when hearing [si] stimuli rather than buzzes, but type of sound hardly affected responses to rises, which could have otherwise been evidence that responses related to native-language prosody. Listeners perceived 375ms or 500ms dynamic stimuli as longer than level more often when hearing [si] stimuli rather than buzzes, whereas type of sound hardly affected responses to 250ms stimuli. When hearing buzzes,

listeners more often perceived an initial (as opposed to a final) dynamic stimulus as longer than level, whereas when hearing [si] stimuli, the order of dynamic/level stimuli had less effect.

Although listeners may not have treated the [si] stimuli as more specific to their language than the buzzes, they could have been aware of the vocal-tract source of [si] (i.e. that it was produced by the human speech mechanism) and its potentially phonological structure, without relating it to a particular language. In categorical perception research, differential responses to speech and non-speech stimuli have been found in several experiments (reviewed in Repp 1984: 289-303), which included stimuli unrelated to speech (e.g. Eimas 1963) and non-speech analogues of consonant cues like VOT (e.g. Liberman et al. 1961), closure (e.g. Perey and Pisoni 1980) and formant transitions (e.g. Mattingly et al. 1971). Repp (1984: 300, original emphasis) noted some controversies, but concluded that ‘there is no conclusive evidence so far for any significant parallelism in the perception of speech and nonspeech. What seems to matter is not whether the stimuli **are** speech or nonspeech, but how listeners **interpret** (i.e. “hear”) them.’ In other words, we have a speech and non-speech ‘mode’ of listening. However, Kingston et al. (2009) investigated the effect of preceding vowel duration on perceived consonant duration using speech stimuli and non-speech analogues, and found that speech and non-speech perception might not always differ. Likewise, Wang et al. (1976) found that a perceived lengthening effect of dynamic f_0 held to the same extent for speech and non-speech stimuli. The present experiment’s results could be explained as a consequence of listeners hearing the two stimulus types in different modes. Importantly though, a perceived lengthening effect of dynamic f_0 *did* occur more often than chance with non-speech stimuli, but the effect was even greater with [si] stimuli.

3.7.3 Nature of stimuli

The similar previous studies differed in the nature of their stimuli: non-speech sounds, synthetic vowels or nonsense/meaningful words; falls, rises or complex f_0 contours; various duration ranges. These factors differentially affected responses in the present experiment, so could be partly responsible for previous conflicting findings, and we should be cautious in directly comparing previous studies to one another. Direction of dynamic f_0 and stimulus duration had main effects (and interactions with sound type) in their respective ANOVAs and the regression analysis (Table 3-8). §3.7.1 discussed the effect of f_0 direction for falls and rises but not complex contours. Stimuli with complex f_0 were judged longer than level significantly more often than rising (but not falling) stimuli were. Perhaps listeners subconsciously reasoned that for a sound to accommodate two f_0 movements (fall-rise/rise-fall), it was probably longer than the comparison sound with no f_0 movement, but this does not explain why falls should have a similarly large effect as complex f_0 .

If we consider stimulus duration, the dynamic stimulus was judged longer than level significantly more often in 500ms and 375ms pairs than in 250ms pairs. The pitch-carrying part

of [si] stimuli was shorter than in the equivalent buzzes, which could be a factor in the different response pattern. The original citation-form [si] monosyllables were around 500ms. The recordings for a later experiment (chapter 4) suggest that around 375ms is a natural duration for [si] produced at a normal rate within longer utterances. 250ms stimuli probably seemed unnaturally short or rushed; listeners may have responded more randomly than for 375ms or 500ms pairs, if they were less sure about which stimulus was longer out of a pair that were both unusually short. This explanation does not account for buzzes, which have no linguistic reference point. House's (1990) claim may help here: listeners' perceptual sensitivity to dynamic pitch decreases in periods of spectral complexity, a moderate level of which occurs at sound onset due to new spectral information. The shorter the buzz, the faster the f_0 movement, and the greater the initial period of new spectral information beginning the stimulus relative to its entire duration. If the f_0 movement is fast, and during relative spectral complexity, listeners might be more inclined to perceive the pitch as level rather than dynamic; thus the comparison they make is no longer dynamic versus level, but level versus level. The same applies to [si] stimuli, since the f_0 movement began during a period of spectral complexity, i.e. the consonant-vowel transition. This does not explain why 375ms, not 500ms, is the optimum stimulus duration for a perceived lengthening effect of dynamic f_0 .

Stimuli with an early- or late-starting f_0 movement were included to investigate House's (1990) claim that f_0 movements are perceived as dynamic during periods of spectral stability, and as level during periods of spectral change (and possibly new spectral information). If a dynamic f_0 is perceived as level, it might not result in an increased perceived length. A significant interaction was found for f_0 timing, f_0 direction and type of sound. For buzzes, listeners perceived late falls as longer than level significantly more often than early falls, but late rises as longer than level significantly less often than early rises; conversely for [si] stimuli, listeners perceived late falls as longer than level significantly less often than early falls, but late rises as longer than level significantly more often than early rises. Following House (1990), late falls and rises should be perceived as falling and rising respectively, since f_0 movement occurred during spectral stability. Late falls had a lengthening effect more often than late rises (though it was fairly equal for [si] stimuli), which corresponds to the response data for falls compared to rises overall. Conversely, early falls (EFs) and early rises (ERs) should be perceived as low- and high-level stimuli respectively, since f_0 movement occurred during changing or new spectral information, and the perceived height depends on the f_0 value once spectral stability is reached (cf. Rossi 1971, 1978). Consequently listeners would hear two low-level stimuli for pairs with an EF, and a high- and low-level stimulus (either order) for pairs with an ER. This no longer concerns dynamic f_0 , rather f_0 height. For buzzes, high (or ER) increased perceived duration more often than low (or EF), whereas for [si] stimuli, high (or ER) increased perceived duration less often than low (or EF). Now consider the trials comprising two stimuli with physically level f_0 , which

showed that for buzzes, but not [si] monosyllables, high-level stimuli were perceived as longer than low-level ones significantly more often than chance (cf. Yu to appear). For buzzes, high increased perceived duration more often than low, whereas for [si] stimuli, high and low influenced perceived duration similarly. These responses generally accord with those for pairs with an EF or ER. This does not discount the idea that f_0 movements are recoded as levels in spectral instability. Again, perception of linguistic and non-linguistic stimuli were quite different.

Finally, an ordering effect was found, which does not confound the overall results, as the stimuli were counterbalanced. Listeners perceived the dynamic stimulus as longer than level more often if the dynamic stimulus came first rather than second. In psychology, ‘contrast effect’ refers to the perceptual effect (in any sensual modality) of a stimulus being perceived differently if placed next to a dissimilar stimulus compared to if isolated. A classic example is that lukewarm water feels warmer if you had your hand in cold water beforehand, compared to if you just place your hand in lukewarm water. Contrast effects have also been noted in speech perception, e.g. Fox (1985) found that vowel identification differed depending on the surrounding acoustic context. In the present experiment, if the dynamic sound had no preceding context, it often seemed longer than the following sound, whereas if the dynamic sound was contrasted with a preceding level sound, the effect of increased perceived duration was attenuated. It is unclear why these results showed the opposite order effect to those of Rosen (1977b), Lehiste (1976) and Pisoni (1976). Responses to the control stimuli (two level sounds) suggest that the first stimulus may generally be perceived as longer, regardless of f_0 contour, as there was a first stimulus response bias in the buzzes (all language groups) and [si] stimuli with SGs but not (S)Frs. In an experiment on prominence perception, Kohler (2008) presented German listeners with disyllabic /baba/ stimuli that had various durational and tonal manipulations; in interpreting the results, he suggested that listeners would expect, but not hear, a longer second syllable in stimuli comprising identical syllables, causing the first to seem more prominent.

3.8 Conclusion

This experiment investigated whether dynamic f_0 influences perceived duration. A previously reported perceived lengthening effect of dynamic f_0 has been observed in three listener groups whose native languages are prosodically different. This effect was greater for linguistically meaningful than meaningless stimuli, which could be evidence that listeners perceived speech and non-speech sounds in different ways; still, the perceptual interdependence of f_0 and duration was evident for both sound types. Although this interdependence of f_0 and duration is independent of native language here, we would need to test more language groups with the same experiment design before we could claim that this is universal in speech perception, mainly because this was not found in previous studies with various languages. There might be an underlying cause, other than tonal and durational properties being interdependent

cues to prominence or vowel-length in a language (Lehiste 1976, Lehnert-LeHouillier 2007), that leads to cross-linguistically different responses; the effect of this cause might not have manifested itself in this experiment's data. It would be interesting (but beyond the scope of this thesis) to use these stimuli to investigate speakers of a tone language (e.g. various dialects of Chinese) in which f_0 movement cues phonological contrasts. Lehnert-LeHouillier (2007) found no perceived lengthening effect of dynamic f_0 with Thai speakers.

Most importantly, this finding that f_0 and duration are perceptually interdependent has a major implication for rhythm research: we should investigate f_0 , as well as duration, because the rhythm of a language which tends to use f_0 dynamism within syllables may be perceived differently from the rhythm of a language in which f_0 changes minimally within syllables. This thesis presents a series of experiments progressing from psychoacoustic/psychophonic to increasingly linguistic tasks. The finding of a perceived lengthening effect of dynamic f_0 at this first stage has laid the basis, but rhythm concerns longer domains than syllable pairs. To observe whether the interdependence of f_0 and duration is not simply a psychoacoustic phenomenon, it is now worth investigating in a more linguistic context. Although [si] was chosen because it corresponds to a frequently used word in the languages investigated, this experiment was unlike natural speech, since these [si] pairs had no context. The following chapter reports an experiment concerning f_0 and duration in rhythmic-group perception.

The interdependence of f0 and duration in rhythmic-group perception

4.1 Summary

The previous experiment found a perceptual interdependence of duration and f0, when the stimuli were pairs of buzzes and monosyllables. The experiment reported here investigates whether dynamic f0 and increased duration are interdependent perceptual cues when listeners are given a more linguistic context, and if so, whether this depends on native language. The stimuli are various digit/letter series, as occur in everyday speech (e.g. telephone numbers, postal codes and serial numbers). The results demonstrate that rising f0 and increased duration are interdependent cues to perceived rhythmic groups within digit/letter series, and that the relative significance of each cue depends on native language. This finding suggests that we need to investigate whether and how duration and f0 are interdependent in the perceived rhythm of linguistically more complex utterances.

4.2 Prosodic groups

The idea that continuous speech comprises definable groups has existed for a long time; Halliday (1960) was the first to explicitly discuss that these groups form a hierarchical structure (Ladd 1996: 237). Since the 1980s, Prosodic Phonology (e.g. Nespor and Vogel 1986, Selkirk 1984) and the Autosegmental-Metrical theory of intonation (e.g. Pierrehumbert 1980) have made this ‘prosodic hierarchy’ concept popular (for a review, see Shattuck-Hufnagel and Turk 1996). There is no general agreement on the number of prosodic levels that should be distinguished (Carlson et al. 2002), and different-sized prosodic groups have been given various labels (see Table 4-1). Often the term(s) that a particular author uses reflects the area of prosody on which they work (cf. Post 2000: 8).

Intonation-related	Rhythm-related	Other
<i>Intonation(al) Phrase/Group/Unit</i>	<i>Rhythmic Unit</i>	<i>Phonological Phrase</i>
<ul style="list-style-type: none"> • Selkirk (1984) • Nespor and Vogel (1986) • Beckman and Pierrehumbert (1986) • Cruttenden (1986) • Hirst (1998) 	<ul style="list-style-type: none"> • Pike (1945) • Guaitella (1999) 	<ul style="list-style-type: none"> • Selkirk (1984) • Nespor and Vogel (1986)
<i>Tone Group/Unit</i>	<i>Rhythmic Group</i>	<i>Accentual Phrase</i>
<ul style="list-style-type: none"> • Palmer (1922) • Halliday (1970) • Gussenhoven (1984) • Crystal (1969) 	<ul style="list-style-type: none"> • Dahan (1996) • Delattre (1966a) • Delais-Roussarie et al. (2002) • Tranel (1987) 	<ul style="list-style-type: none"> • Jun and Fougeron (2000) • Beckman and Pierrehumbert (1986)
<i>Tune</i>		<i>Breath Group</i>
<ul style="list-style-type: none"> • Armstrong and Ward (1926) • Kingdon (1958) 		<ul style="list-style-type: none"> • Pulgram (1965) • Vaissière (1991a, 1991b)
		<i>Sense Group</i>
		<ul style="list-style-type: none"> • Vanderslice and Ladefoged (1972)

Table 4-1 – Examples of labels for variously-sized prosodic groups

Although prosodic groups are widely accepted phenomena, it is often unclear how to identify them in spontaneous speech data (Cruttenden 1986, Crystal 1969, Ladd 1996). In Prosodic Phonology, groups are based on syntactic structure, though arguably they are difficult to identify according to syntax (Ladd 1996: 235-36). (Cutler et al. (1997: 159-71) review studies concerned with the relationship between prosody and syntax.) In Autosegmental-Metrical theory, groups are based on phonetic features, e.g. intonation and boundary phenomena (for a summary, see Jun 1998). Carlson et al. (2002) suggested that groups are difficult to identify phonetically because we need more research on how groups are *perceptually* cued by f₀ movement, pre-boundary lengthening, voice-quality changes etc. Similarly, Guaitella (1999) argued that groups are perceptual phenomena, so their form is unlike that of something produced as a group.

Nevertheless, there is some perceptual evidence for prosodic groups, as the following experiments exemplify. Early psychologists investigating (non-speech) rhythm identified the effect of certain acoustic properties on grouping perception. Bolton (1894) found that in a series of clicks, those with stronger intensity were perceived as group-initial, and those with longer duration as group-final. In similar experiments, Woodrow (1911) found that intensity and duration affected grouping as Bolton (1894) had observed, whereas higher pitch had neither a group-final nor group-initial effect. Fraisse (1982: 155-64) summarised psychological research along these lines.

With speech stimuli, Hay and Diehl (1999, 2007) found that English and French speakers perceived rhythmic groups within series of /ga/ syllables; again, increased duration had a group-final effect and stronger intensity a group-initial effect. De Rooij (1976) and Bouwhuis et al. (1978) constructed series of seven /da/ monosyllables with different f₀ contours and certain lengthened vowels. For each series, listeners had to choose the best-fitting sentence from a list of syntactically varied Dutch sentences. Results showed that group-final lengthening, and lengthening plus dynamic pitch, but not dynamic pitch alone, were effective cues to within-sentence prosodic groups. De Pijper and Sanderman (1994, see also Collier 1993) and Swerts (1997) found that untrained listeners perceived boundaries of different prosodic strength (i.e. different groups in the hierarchy) in sentences; inter-subject agreement correlated positively with boundary strength, and strength ratings were higher the more phonetic cues were present (e.g. f₀ movement, pre-boundary lengthening, pauses). Carlson and Swerts (2003) demonstrated that when these phonetic cues were present, listeners could predict upcoming boundaries. Cutler et al. (1997: 162) summarised experiments that showed which cues listeners used to locate prosodic-group boundaries in speech: durational (e.g. Lehiste 1973), tonal (e.g. Cooper and Sorenson 1977) or amplitude-related properties (e.g. Scholes 1971). Several experiments have shown that listeners recalled a word/digit series better if it comprised groups with boundaries marked by phonetic cues like pauses and f₀ movements (references in Boucher 2006: e.g. Crowder 1982,

Frankish 1985, Frankish 1989, Reeves et al. 2000, Wickelgren 1967). Reeves et al. (2000) concluded that their listeners, who were not instructed to use grouping strategies, spontaneously used prosodic cues to group stimuli and recall information. Boucher (2006) suggested that memory contributes to the nature of prosodic structures, since the number of syllables in a group at which listeners' recall was greatest corresponded to the most likely number of syllables in a stress-group produced by the same subjects.

The present experiment was inspired by House's (1990) research. He recorded a Swedish speaker repeating the word [fɛ̃m] ('five'); one token was selected, multiplied, and the f0 contour was variously manipulated on different replicates, which were then spliced together to form stimuli of five fives ('55555') with various tonal configurations. Listeners indicated whether they heard 55-555 or 555-55. Some listeners listened to fives with fall-rise or rise-fall contours, others listened to fives with falls, rises or level contours, which tested how dynamic f0 and mean f0 level contributed to their perception of group boundaries. The strongest group-boundary cues were a clear f0 level difference between the final five of the first group and the initial five of the second group, and a 100Hz fall on the final five of the first group which tended to be lower than the initial five of the second group. The tonal similarity of certain fives may also have contributed to them being perceived as a group (House 1990: 95). Duration could conceivably contribute as a group-boundary cue in this context, though House did not investigate this. During stimulus preparation, House (personal communication 2008) thought that the 100Hz falls (and rises), which turned out to be particularly effective boundary cues, sounded longer than the other f0 contours, which gave him a clear rhythmic-grouping percept.

An obvious next step, which the present experiment takes, is to manipulate f0 *and* duration in such stimuli, to test whether these are interdependent cues to rhythmic groups within five-syllable series. (The term 'rhythmic group' was chosen primarily because this thesis concerns rhythm; terms referring to prominence were avoided, since prominence differs between (S)Fr and SG.) The experiments discussed above provide evidence for the psychological reality of within-utterance groups, the initial/final boundaries of which may be cued by durational, tonal or other properties. The present experiment adds to the relatively limited perceptual data on prosodic-group boundary cues, in particular by investigating the influence of listeners' native language, which is under-researched (Cutler et al. 1997). We should bear in mind that, in general, prosodic-group perception might not concern phonetic cues to boundaries alone, but also groups' internal structural coherence. For example, a recurring pattern of 'high-low-high' pitch may lead to a perceived boundary between the final high of one group and the initial high of the next, even though no clear acoustic difference is observable between these highs.

4.3 Hypothesis

Experiment 1 (chapter 3) demonstrated that f_0 and duration are interdependent in a psychoacoustic/psychophonetic task. Previous experiments on perceived grouping demonstrated that durational and/or tonal cues signal prosodic-group boundaries, which suggests that durational and tonal cues could be interdependent in the perception of rhythmic groups in this experiment. It is predicted that listeners will locate rhythmic-group boundaries using length and pitch cues as follows:

- when only one cue is available, the number of responses indicating that increased duration or dynamic f_0 was the cue that listeners used will be high;
- when both cues are available and accord (i.e. the same syllable has increased duration and dynamic f_0), the number of responses indicating that these cues were used will be even higher;
- when both cues are available and conflict (i.e. one syllable has increased duration and a neighbouring syllable has dynamic f_0 around the possible boundary location), the number of responses indicating that these cues were used will be considerably lower;
- when neither cue is available, responses will be around chance.

No prediction is made concerning native language, mainly because experiment 1 found no evidence that this affected responses, and few perceptual grouping experiments have provided evidence for cross-linguistic variation (except in non-speech, e.g. Iversen et al. 2009). In this more linguistic task, if listeners' use of length and pitch cues turns out to depend on native language, interpretation of the results will address this.

4.4 Subjects

All subjects (Table 4-2) reported normal hearing.

Language		Total	Age (years)		Sex	
Monolingual	SG	36	Range	20–38	Male	6
			Mean	25.7	Female	30
	SFr	38	Range	18–39	Male	15
			Mean	24.1	Female	23
	Fr	36	Range	18–35	Male	14
			Mean	24.4	Female	22
Bilingual		20	Range	17–37	Male	5
			Mean	23.8	Female	15

Table 4-2 – Summary of subjects

If native-language prosody influences a perceptual interdependence of duration and f_0 , subjects who acquired both SG and SFr in infancy (i.e. bilinguals) might behave differently when completing the task in each language, since prosody perception develops early in infancy (see e.g. Nazzi et al. 1998, Nazzi et al. 2000, Nazzi and Ramus 2003, Nazzi et al. 2006). If native language has no effect, all monolinguals and bilinguals should respond similarly. With either outcome, bilinguals provide more insight into how language-universal/-specific the perceptual interdependence of duration and f_0 is; this is why they were tested. It was not viable to collect enough bilingual data for the other experiments in this thesis, because relatively few Swiss people acquire two languages in infancy; most acquire one native language depending on the community where they grew up (see chapter 2). During three fieldwork trips, the number of bilingual volunteers available was only sufficient for one experiment. This one was most appropriate because it took the shortest time, so bilinguals could do it in both languages within a reasonable time period.

A simple definition of ‘monolingual’ and ‘bilingual’ does not exist (for discussion, see Galloway 2007). In this experiment, distinct monolingual and bilingual groups were defined by three factors: ambient language in childhood, and age and method of second-language acquisition. For monolinguals, each subject grew up in one country with one ambient language in childhood; some had temporarily lived abroad in adulthood, e.g. a gap year after schooling. None had started to learn a second language before compulsory second-language classes in school (starting age 9-11). SFr speakers learn standard German at school (see Galloway 2007), so most speak no SG; a few SFr subjects had learned some SG from living in Zürich in adulthood. The locations where the monolinguals were recruited and tested, and other details like their occupations and regional origins, were as for SG/SFr/Fr subjects in experiment 1 (chapter 3). Table 4-3 summarises the bilinguals’ language acquisition backgrounds. Those marked * (nine) had parents who each spoke a different language to them from birth. Those marked ** (eleven) were brought up in an area where the ambient language was different from that spoken at home by parents, so they were immersed in the other language from an early age through schooling and friendships with local children. The bilinguals had to rate how native-like their spoken language and comprehension currently are in both languages; all scored themselves highly. They were tested in Zürich University Phonetics Laboratory’s sound-attenuated booth, or in a quiet room in Neuchâtel University.

	Parents' native language(s)		Place of birth	Place(s) lived in during childhood	Age & method of acquisition (H: home; F: friends) (explanation below)	Current ability (self-scored, 0-5: 0 = nothing, 5 = native-like)			
	Mother	Father				SG		SFr	
						Speaking	Comprehension	Speaking	Comprehension
1 **	SG, SFr	SFr	Morges (SFr)	Baden, Endingen (SG)	SFr: 0, H SG: 4, F	5	5	4-5	5
2 **	SFr	SG	Thalwil (SG)	Hausen am Albis (SG)	SFr: 0, H SG: 2-3, F	5	5	5	5
3 *	SFr	SG	Basel (SG)	Basel (SG)	SFr: 0, H SG: 0, H	5	5	4	4
4 *	SFr	SG, SFr	Payerne (SFr)	Zürich (SG)	SFr: 0, H SG: 0, H then F	5	5	5	5
5 **	SFr	SFr	Basel (SG)	Basel (SG)	SFr: 0, H SG: 4, F	5	5	4	4
6 *	SG	SFr, Arabic	Geneva (SFr)	Geneva (SFr)	SFr: 0, H then F SG: 0, H	5	5	5	5
7 **	SFr	Italian	Zürich (SG)	Zürich (SG)	SFr: 0, H SG: 3, F	5	5	5	5
8 **	SG	SG, SFr	Sierre (SFr)	Sierre (SFr)	SFr: 3, F SG: 0, H	5	5	4-5	4-5
9 **	German	SG	Geneva (SFr)	Geneva (SFr)	SFr: 3, F SG: 0, H	4	4	5	5
10 **	SFr	SG	Zürich (SG)	Zürich (SG)	SFr: 0, H SG: 4, F	4.75	5	4	4.5
11 *	SG	Polish (SFr to child)	Zürich (SG)	Montreux (SFr)	SFr: 0, H then F SG: 0, H	5	5	5	5
12 **	SFr	SFr	Geneva (SFr)	Basel (SG)	SFr: 0, H SG: 3, F	5	5	5	5
13 *	SG	SFr	Aigle (SFr)	Villeneuve, Thalwil (SFr, SG)	SFr: 0, H SG: 0, H	5	5	5	5
14 *	SG	SFr	Zürich (SG)	Zürich (SG)	SFr: 0, H SG: 0, H then F	5	5	5	5
15 **	SFr	SFr	Sion (SFr)	Winterthur (SG)	SFr: 0, H SG: 4, F	4	5	5	5
16 *	SG	SFr	Zürich (SG)	Zürich (SG)	SFr: 0, H SG: 0, H then F	5	5	5	5
17 *	SG	SFr	Munich (German)	Biel/Bienne (SG/SFr)	SFr: 0, H then F SG: 0, H then F	5	5	5	5
18 **	Arabic (SFr to child)	SFr	St-Aubin (SFr)	Rohr (SG)	SFr: 0, H SG: 3, F	5	5	5	5
19 *	SG	SFr	Biel/Bienne (SG/SFr)	Brügg (SG), Biel/Bienne (SG/SFr)	SFr: 0, H then F SG: 0, H then F	4	5	5	5
20 **	SG	SG	Basel (SG)	Epiquerez (SFr)	SFr: 4, F SG: 0, H	4	5	5	5

- 'Friends' (F) means they started acquiring the language when they went to kindergarten or started socialising with other children of that language.
- '0' years indicates: from birth onwards, whenever the child began first-language(s) acquisition.

Table 4-3 – Summary of bilinguals' language backgrounds

4.5 Method

A forced-choice AB task was designed and run in *Praat* (Boersma and Weenik 2008-2009: version 5.0.21). Two sets of stimuli, one SG one Fr, were carefully designed to be cross-linguistically comparable and equally appropriate for each language group (this is a challenge of cross-linguistic research; see Beddor and Gottfried 1995). Each monolingual group heard stimuli in their native language (Fr for SFr subjects, since the words used do not show between-variety variation). Bilinguals completed the same task in each language. The stimuli were sequences of five digits and/or letters, in which duration and f_0 were systematically manipulated on the second and/or third syllable. Each sequence comprised syllables that were taken individually from naturally produced sequences, then resynthesised with specific durations and f_0 contours, and then spliced together. This procedure is now described in detail.

4.5.1 Recordings

One native SG speaker (Zürich dialect) and one native Fr speaker (from Rennes, France) were recorded; neither was phonetically trained. They were each asked to read, at a comfortable rate, a list of digit and letter sequences in the format:

(X X X) (X X)

(X X) (X X X)

where X was a monosyllabic digit/letter, all Xs

in each sequence were identical, and each grouping (3+2, 2+3) occurred once for all single digits (1-9) and all monosyllabic letters of the alphabet (A-Z). The order of digits and letters down the reading sheet was randomised. There were two reasons for recording sequences in this format, rather than isolated syllables or ungrouped sequences of five. First, this checked that the speakers understood the notation for grouping, which they did without instruction. One suggested that XXX-XX might be more obvious, as did a few listeners in a pilot of the perceptual test, hence the use of XXX-XX in the perceptual experiment. Second, this ascertained that speakers produced longer syllables with an f_0 rise or fall before each group boundary, and the manipulations later made on the stimuli were based on these duration and f_0 values.

One speaker at a time was recorded in a sound-attenuated studio, using a *Marantz* PMD670 solid-state recorder and a low-noise condenser *Sennheiser* MKH40P48 microphone with a cardioid frequency response. The recording mode was set to 16 bit linear PCM, with a 44.1 kHz sample rate. The files were saved as .wav format, then transferred onto a *MacBook* (Mac OS X.4) via a USB cable, and displayed in *Praat*. The recordings were visually and auditorily inspected to choose the most appropriate and cross-linguistically equivalent individual syllables to make up the stimuli. These stimuli (unlike House's (1990)) mostly comprised syllables with various segmental structures (henceforth 'varied stimuli'), so that listeners did not hear a constant repetition of one sound, thus ensuring that they heard the stimuli as meaningful digit/letter sequences and did not become bored. Some stimuli, also created by splicing together individual

syllables, had segmentally identical syllables (henceforth ‘identical stimuli’), which were included to compare responses with those to the varied stimuli (see Table 4-4).

	SG		Fr		Pattern (see explanation below)
Identical	BBBBB	/be/ (x5)	PPPPP	/pe/ (x5)	■ ■ ■ ■ ■
	22222	/tsvai/ (x5)	33333	/tɕwa/ (x5)	■ ■ ■ ■ ■
Varied	3PTP3	/dry.pe.te.pe.dry/	2BDB2	/dø.be.de.be.dø/	■ ◆ ○ ◆ ■
	5Z4J6	/foif.tset.fier.jɔt.saxs/	75Z46	/set.sɛk.zed.katɕ.sis/	■ ◆ ○ □ ☆
	H2H2H	/ha.tsvai.ha.tsvai.ha/	C3C3C	/se.tɕwa.se.tɕwa.se/	■ ◆ ■ ◆ ■
	S888F	/ɛs.ɥt.ɥt.ɥt.ɛf/	S888F	/ɛs.ɥit.ɥit.ɥit.ɛf/	■ ◆ ◆ ◆ ○

Table 4-4 – Syllable sequences chosen for stimuli

There were various reasons for choosing these digits/letters. First, all had a sonorant nucleus, and a voiceless onset and/or coda, which ensured a between-syllable break of periodicity, so f_0 manipulations would have clear start and end points. The plosives /b d/ are unvoiced lenis in SG (see Fleischer and Schmid 2006), and usually voiced in Fr but still have a burst transient disrupting periodicity. Second, the SG and Fr sequences’ segmental structure had to be cross-linguistically as similar as possible. The main issue was choosing syllables in which the voiceless and sonorant parts were durationally similar in each language, so that these parts could be manipulated to the same duration for equivalent stimuli in each language. For example, the SG and Fr plosives /p t b d/ had similar durations, as did the SG and Fr fricatives /h f s/, the SG affricate /ts/ and Fr plosive+fricative /tɕ/, the SG and Fr vowels /e ɛ a/, and the SG diphthong /ai/ and Fr approximant+vowel /wa/. Third, the ‘5Z4J6’/‘75Z46’ stimuli were designed to have five segmentally different syllables, which were different from all others in the stimuli. The varied stimuli, as well as presenting a changing signal, tested whether listeners tended to group certain syllable patterns (see the right-most column of Table 4-4) using criteria other than prosodic cues. For instance, a listener might always hear 3PTP3 as 3P-TP3, regardless of tonal and durational properties, if that listener thinks letters sound better together group-initially for some reason like his/her postal-code system is similar.

To create the sequences in Table 4-4, fourteen individual syllables per language were needed from the recordings. Figures 4-1 and 4-2 present measurement data averaged across the fourteen sequences which contained these syllables.

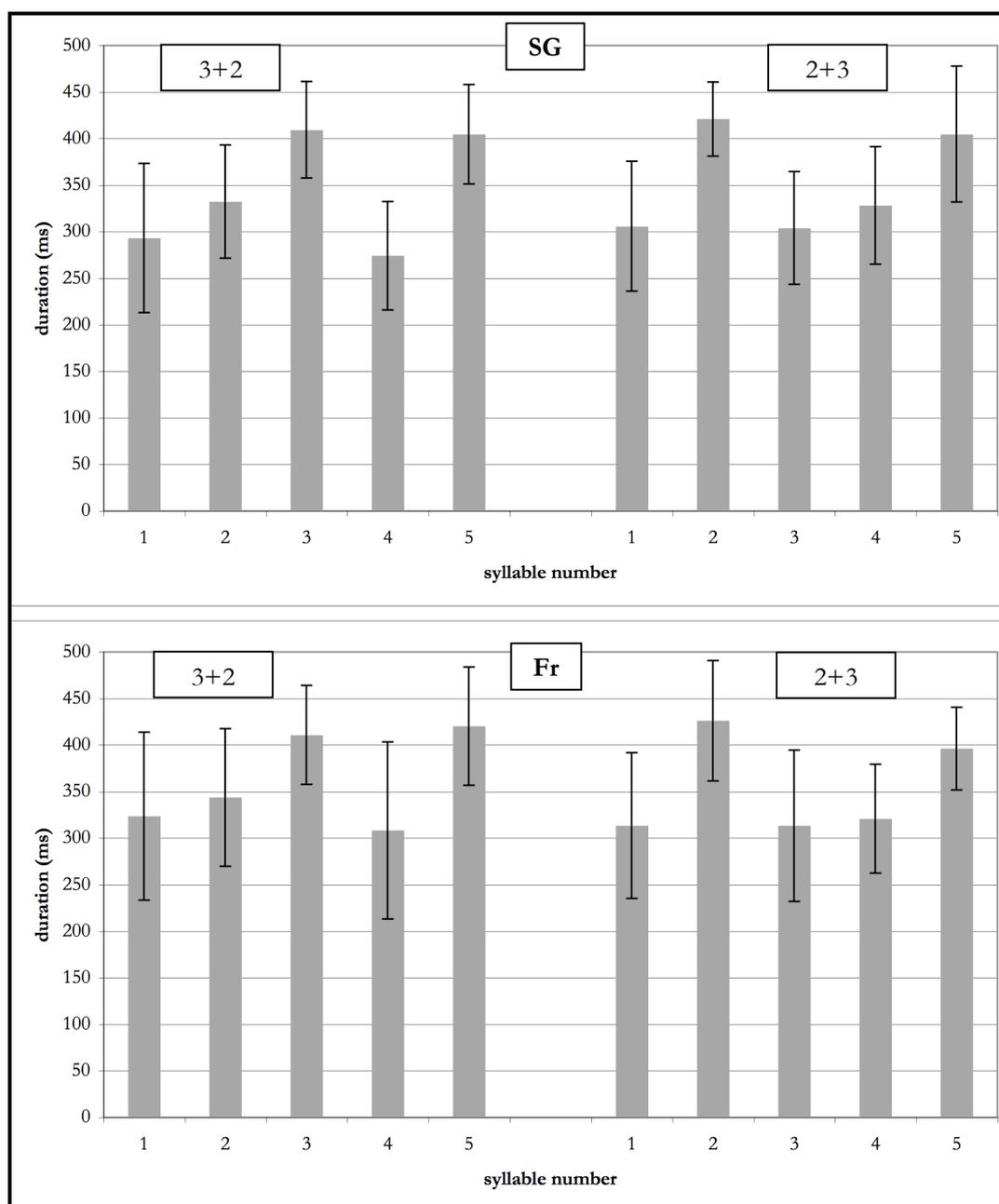


Figure 4-1 – For each syllable (1-5): mean duration across 14 sequences (error bars: ± 1 standard deviation)

The duration of group-initial plosives was measured from the release burst, since closure duration was unknown. We see that the mean duration of respectively positioned syllables was similar across languages. The standard deviations reflect the variation that resulted from different syllables' segmental structure.

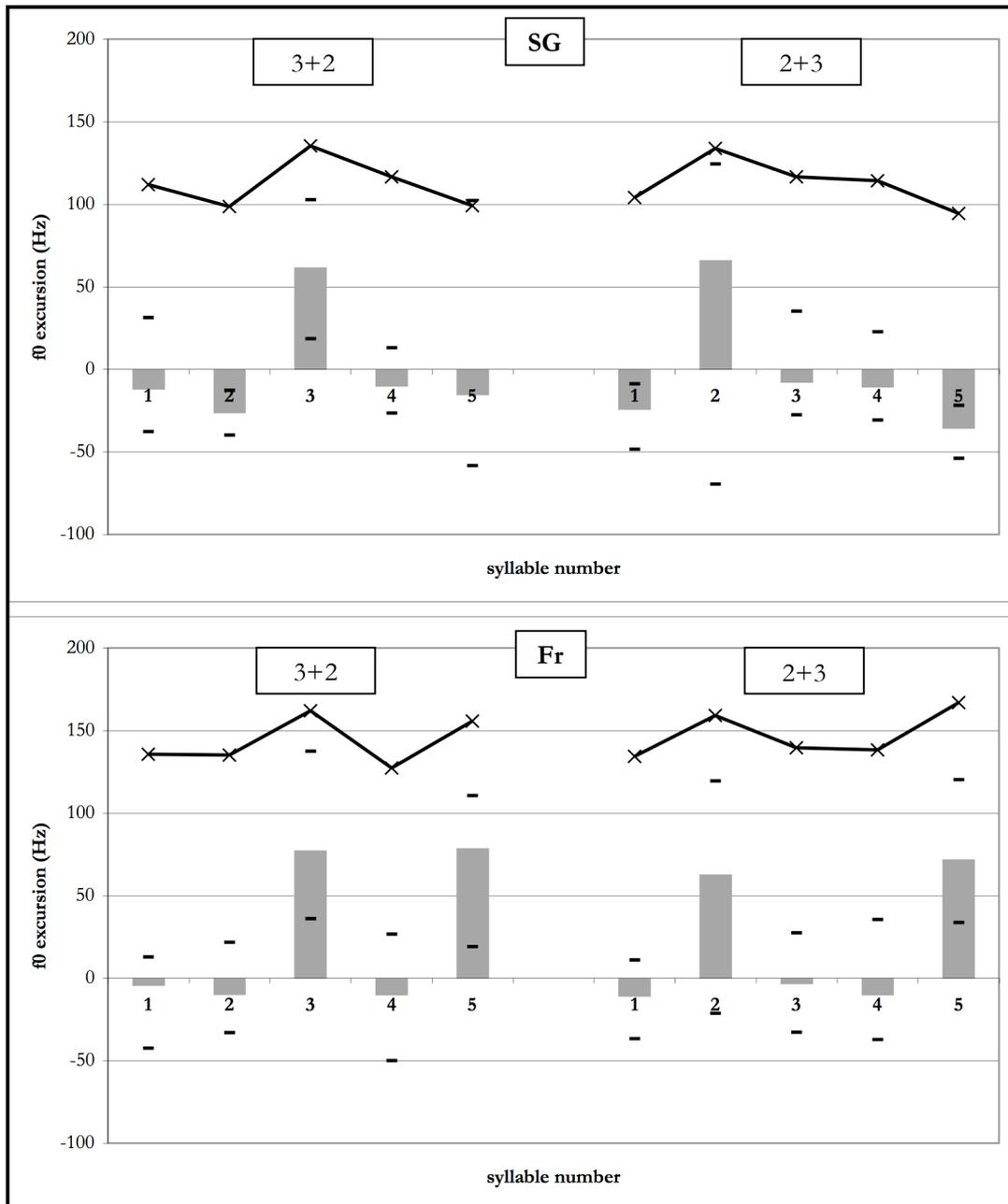


Figure 4-2 – For each syllable (1-5): across 14 sequences, mean f0 excursion (grey columns); minimum/maximum f0 excursion (black bars); mean average f0 (black line)

More inter-speaker variation is noticeable for f0 than duration. The SG speaker's voice was slightly lower in pitch than the Fr speaker's. On the sequence-final syllable, the SG produced mainly falls, whereas the Fr produced only rises, except a fall on the last sequence of each page of the list (though neither page-final sequence was needed for stimulus preparation), which suggests that the Fr rises were a continuation intonation pattern resulting from list reading.

4.5.2 Stimulus preparation

After analysis of the recordings, which led to decisions on which digits/letters to use in stimuli and provided data on durational and tonal properties of rhythmic groupings, the stimuli

were created in *Praat*. In outline, this comprised the following steps, which are then detailed below.

1. One token of each of the digits/letters required for each language (i.e. B, P, Z, 2, 3, 5 etc.) was selected.
2. Acoustic properties were adjusted to make all tokens equivalent within and between languages (i.e. 'base' syllables).
3. Each base syllable was multiplied, and duration and/or f_0 was manipulated on the replicates.
4. These manipulated syllables were concatenated into five-syllable sequences.

4.5.2.1 Token selection

For each syllable required, there were ten possible tokens: five per recorded sequence, one grouped 3+2 and one 2+3. The most appropriate token was selected using various criteria, through visual inspection of the spectrogram and waveform, and simultaneous auditory observation. First, any syllables with non-modal voicing (e.g. creak) were eliminated, since these could be problematic during resynthesis. Then only non-group-final syllables (i.e. 1, 2, 4 in 3+2 sequences and 1, 3, 4 in 2+3 sequences) were considered, because the 'base' syllable to be made from each selected token would sit in these positions in stimuli and have a level f_0 and normal (i.e. not increased) duration. Less manipulation was required by selecting a token already not very deviant from these properties, compared to a group-final syllable on which durational and tonal boundary effects were present. The aim was to manipulate as little as possible, to minimise artificiality. Ultimately, syllable 2 in 3+2 sequences or syllable 4 in 2+3 sequences was the preferred choice for all tokens, because several of the required syllables began with plosives, for which only group-medial tokens were suitable, since the closure duration of post-silence plosives was unknown.

The overall balance between minimising artificiality of stimuli and maintaining strict control over prosodic manipulations required compromises. The use of real speech increased naturalness, particularly since the stimuli were based on the prosodic properties of the produced utterances. Yet to systematically vary duration and f_0 , stylisation and resynthesis were needed, which increased artificiality. An inevitable consequence of taking syllables from continuous speech was the presence of formant transitions at syllable boundaries, which could have decreased naturalness when syllables were removed from one sequence and concatenated with others. However, the task involved attention to prosody rather than segments, and listeners had a simultaneous visual presentation of the digits/letters, so would not be confused by perhaps inappropriate transitions at syllable boundaries, nor distracted from making decisions based on prosodic cues.

4.5.2.2 ‘Base’ syllables

Once each selected digit/letter token had been extracted from the recording and saved as a separate .wav file, duration, intensity and f0 were altered to make all tokens equivalent. The result was a base syllable for each digit/letter, which could then be subject to further manipulation.

Duration: A *Praat* script was written and run which changed the consonant and vowel duration in each syllable by a certain percentage during resynthesis, to make segment durations identical between languages. Table 4-5 shows the resulting duration (d1) of each resynthesised syllable, which was saved. These durations were chosen as a compromise between the duration of segments in each language’s recordings (§4.5.1). Vowel duration was shorter in CVC syllables than in others. The total durations are comparable to the durations of shorter syllables in Figure 4-1.

SG	Fr	C(C) (ms)	V (ms)	C(C) (ms)	Total (ms)
B, 3, P, T	P, 2, B, D	70	250		320
2, H	3, C	100	250		350
4, 5, 6, J, Z	4, 5, 6, 7, Z	95	175	95	365
8, F, S	8, F, S		250	120	370

Table 4-5 – Base duration (d1) of each syllable (V included the approximants in Fr /ʒit/ ‘8’ and /tʁwa/ ‘3’, and the short voiced trill in SG /dry/ ‘3’)

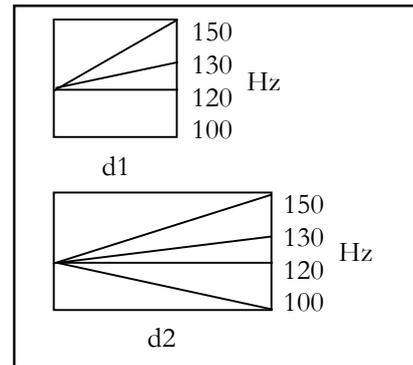
Intensity: In each syllable, peak intensity was manipulated to 83dB, using *Praat’s IntensityTier* function, which allows a relative dB value to be specified and multiplied with a sound file’s existing intensity during resynthesis.

f0: In each syllable, f0 was manipulated to a level 120Hz, using *Praat’s PitchTier* function, which allows an f0 contour to be specified and replace a sound file’s existing f0 during resynthesis. 120Hz was a compromise between each speaker’s mean pitch (Figure 4-2).

4.5.2.3 Duration/f0 manipulations

At this point, the syllables were in their base form (duration=320-370ms (Table 4-5), intensity=83dB, f0=120Hz). Each base syllable was replicated six times; each replicate then underwent a durational and/or tonal manipulation that would act as a potential rhythmic-group cue in the final stimuli. The six manipulations are represented schematically in Figure 4-3 by the diagonal and horizontal lines within the boxes; 120Hz at d1 is the base form, and d2 refers to increased duration (see Table 4-6). Duration was manipulated first (where needed), by running a *Praat* script written for this purpose; f0 was then manipulated using the *PitchTier* function (see §4.5.2.2).

Figure 4-3 – Six manipulations (two durations, four possible f₀ excursions) and the base form (d1, 120Hz)



SG	Fr	C(C) (ms)	V (ms)	C(C) (ms)	Total (ms)
B, 3, P, T	P, 2, B, D	70	375		445
2, H	3, C	100	375		475
4, 5, 6, J, Z	4, 5, 6, 7, Z	125	263	125	513
8, F, S	8, F, S		375	160	535

Table 4-6 – Duration of each syllable after lengthening (d2)

In the recordings for both languages, vowels lengthened more than consonants group-finally. Therefore, for the manipulations, vowel durations at d2 were a 50% increase on d1, whereas consonant durations were kept equal at d1 and d2, except the syllables without an onset consonant. In these, if the coda consonant was kept at 120ms, a 375ms vowel sounded odd (compared to the recordings), so the coda was increased until it sounded sufficiently acceptable (160ms). The total durations are comparable to the higher durations of group-final syllables in Figure 4-1.

An f₀ fall was only implemented on d2 (longer) syllables, for the reason given shortly (§4.5.2.4). Rises with two different f₀ excursions (10Hz, 30Hz) were created to test whether listeners responded differently when the magnitude of f₀ excursion around the rhythmic-group boundary was like a pitch-accent or a microfluctuation (as will be shown in Figure 4-4). Originally a 50Hz rise (around the speakers' mean excursion on group-final syllables: Figure 4-2) was implemented instead of 30Hz, but the 50Hz rise sounded extraordinarily conspicuous (compared to the recordings). Perhaps the perceived difference between rise and 'non-rise' was greater when the 'non-rises' were completely level (in the stimuli), than when the 'non-rises' had minor excursions (in the recordings). 30Hz sounded more comparable to the recordings, and a pilot of the perceptual task demonstrated that listeners perceived the 30Hz rise differently from the 10Hz rise and comparably to the lengthening manipulation.

4.5.2.4 Concatenation of syllables

The sequences in Figure 4-4 were created for each digit/letter pattern (Table 4-4) by splicing together, without pauses, the appropriate syllables created in the previous stage. Bear in mind that listeners had to say whether they heard a 3+2 or 2+3 grouping.

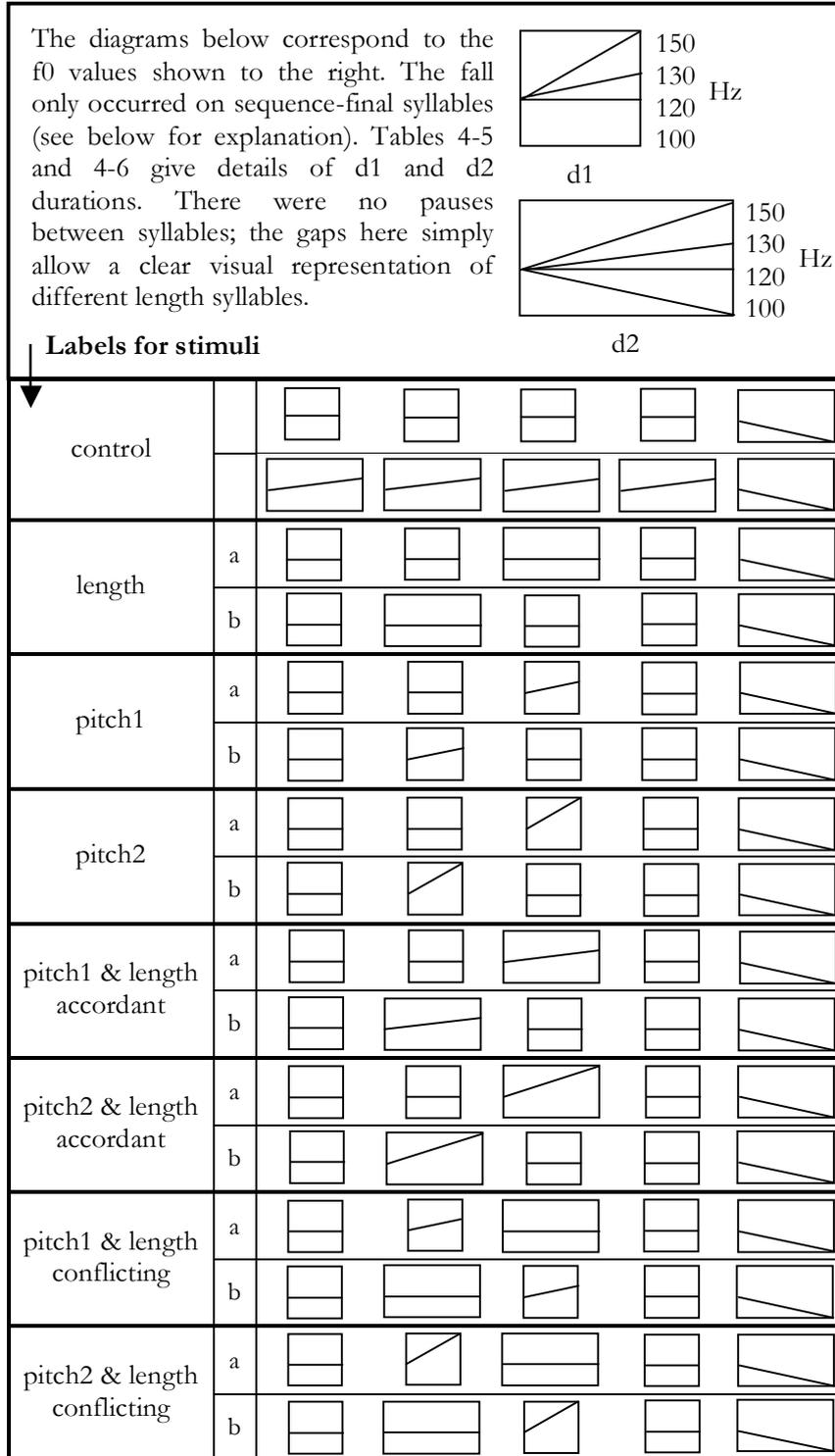


Figure 4-4 – Stimuli: sequences of durationally/tonally manipulated syllables spliced together

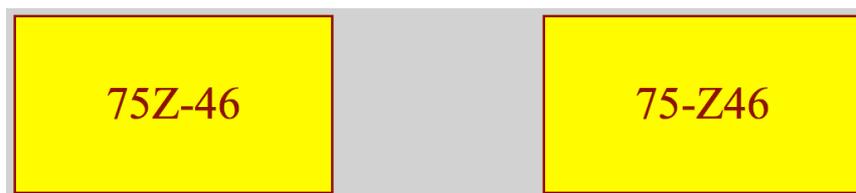
Some commentary on Figure 4-4 follows.

- The acoustic properties manipulated were duration and f_0 , whereas the stimulus labels include ‘length’ and ‘pitch’, which refer to perceptual cues and distinguish these from acoustic manipulations.
- The control stimuli test whether a bias for 3+2 grouping occurs when neither cue is available (cf. House 1990). The remaining stimuli were counterbalanced to cancel out potential bias; for each cue condition, a pair of stimuli were created (‘a’ and ‘b’) which are identical except that the second and third syllables are sequentially opposite.
- The only dynamic f_0 tested here is a rise, rather than various directions like experiment 1 (chapter 3), primarily because in the recordings almost all syllables immediately before the first rhythmic-group boundary had rises.
- To increase naturalness, each sequence ends in a long fall. Both speakers produced long sequence-final falls (Figures 4-1 and 4-2). Although the Fr speaker also produced sequence-final rises (§4.5.1), these seemed to be a continuation intonation pattern from list reading. This was inappropriate for the stimuli, in case listeners got confused by thinking that their grouping judgement should include speech that was potentially going to come after the five-syllable utterance.
- The ‘pitch & length accordant’ and ‘pitch & length conflicting’ stimuli test whether pitch and length cues are interdependent in rhythmic-group perception (see §4.3). In ‘accordant’ stimuli, the same syllable has increased duration and dynamic f_0 , whereas in ‘conflicting’ stimuli, one syllable has increased duration and a neighbouring syllable has dynamic f_0 around the possible boundary location. Conflicting cues could have been placed on the same syllable (e.g. dynamic f_0 and shorter duration); this alternative was investigated during stimulus preparation. However, this was rejected since the resulting sequences sounded odd, and listeners might have perceived a group boundary after such an anomalous ‘conflicting’ syllable because it was an unexpected break from normality in the signal rather than because its duration/ f_0 were cues. Conflicting cues on successive syllables could also sound unusual, but listeners must still decide between two anomalous syllables which one is group-final. This eliminates the potential confound that listeners always perceive a boundary after one anomalous syllable due to non-prosodic factors, and more clearly concerns the interaction of duration and f_0 .

4.5.3 Procedure

Subjects sat at a *MacBook* laptop (Mac OS X.4) and listened through binaural *Sennheiser* HD555 headphones. The experiment was scripted and run in *Praat*. There were two identical versions, one with SG stimuli and German on-screen text, one with Fr stimuli and on-screen text. The procedure for monolinguals was as follows. Before testing began, subjects read instructions in their native language (appendix 8.2.1; German for SGs, Fr for (S)Frs), and were given chance to ask for clarification. For each trial, the following statement appeared on screen:

‘These digits/letters could be grouped as...’. The task was to click one of the two on-screen buttons labelled with a 3+2 or 2+3 grouping, e.g.



Half the subjects saw 3+2 on the left and 2+3 on the right; the other half saw the reverse. Each trial began after subjects had clicked to begin the experiment, or to register their response to the previous trial. After this click came 1.5 seconds of silence before the stimulus. Only one listening per trial was allowed, and response time was unlimited. There were 102 trials (Table 4-7, including practice trials) with a break after every sixteen. Subjects could also ask for clarification after the practice session, though none did. The whole experiment lasted approximately fifteen minutes.

<i>Practice</i>	a range of stimuli (one of each digit/letter pattern, with various manipulations)			6
<i>Main</i>	6 digit/letter patterns (see Table 4-4)	× 8 cue conditions (see Figure 4-4)	× 2 orders (see Figure 4-4)	96
	Total			102

Table 4-7 – Number of trials

For bilinguals, the procedure was very similar, except that they completed the task in each language with a break in between. They had a practice session for each language, to allow them to adjust to the new speaker. Half listened to SG first, and half to Fr first; within these two groups, half saw 3+2 on the left of the screen, and half saw 3+2 on the right. Subjects had been assigned to one of these four groups before they arrived. They read the instructions and conversed with the experimenter (e.g. to ask for clarification) in the language that they listened to first.

4.5.4 Analysis

Responses were recorded in *Praat*, transferred to *Excel*, and sorted according to the variables in Table 4-8.

Variable	Level	
native language(s)	SG	
	SFr	
	Fr	
	BiSG	
	BiSFr	
cue(s)	control	
	length	
	pitch1	
	pitch2	
	pitch1 & length accordant	
	pitch2 & length accordant	
	pitch1 & length conflicting	
	pitch2 & length conflicting	
digit/letter pattern	BBBBB	PPPPP
	22222	33333
	3PTP3	2BDB2
	5Z4J6	75Z46
	H2H2H	C3C3C
	S888F	S888F
order	a	
	b	

Table 4-8 – Variables in experiment design

Data analysis consisted of four stages.

1. Responses to control stimuli were analysed, to ascertain whether a bias for 3+2 or 2+3 grouping occurred (cf. House 1990).
2. All variables were included in one logistic regression model using the *R* software environment, to reveal which variables significantly affected responses.
3. The digit/letter pattern variable, which had been included to present subjects with a continuously changing signal, was investigated together with native language(s), using ANOVAs run in *SPSS*.
4. The main analysis concerned the two variables crucial to the experiment's aim: cue(s) and native language(s). The data were averaged over the six digit/letter patterns to avoid a three-way comparison (language(s) \times cue(s) \times pattern) in which any interactions would be complex to interpret with the non-parametric tests which would be required. The averaged data were

suitable for ANOVAs, which were run in *SPSS*. Most data-series were relatively normally distributed according to visual inspection of histograms, though some were significantly non-normal according to Shapiro-Wilk tests and skewness/kurtosis statistics ($p < 0.05$) (Field 2005). Given that ANOVA is robust against violation of the normality assumption (Howell 2007: 316), and that transformations (including square root and arcsine) made little difference to the normality of these data, untransformed data were input to the ANOVAs. No data-series violated the homogeneity of variance assumption according to Levene tests ($p > 0.05$), which was important because ‘heterogeneity of variance and unequal sample sizes do not mix’ (Howell 2007: 316); two more SFrs participated than SGs and Frs, and the monolingual-bilingual comparisons had even larger disparity of sample size.

4.6 Results

4.6.1 Control stimuli

Figure 4-5 displays the responses to control stimuli (two per digit/letter pattern), in which neither f_0 nor duration was manipulated on the second or third syllable.

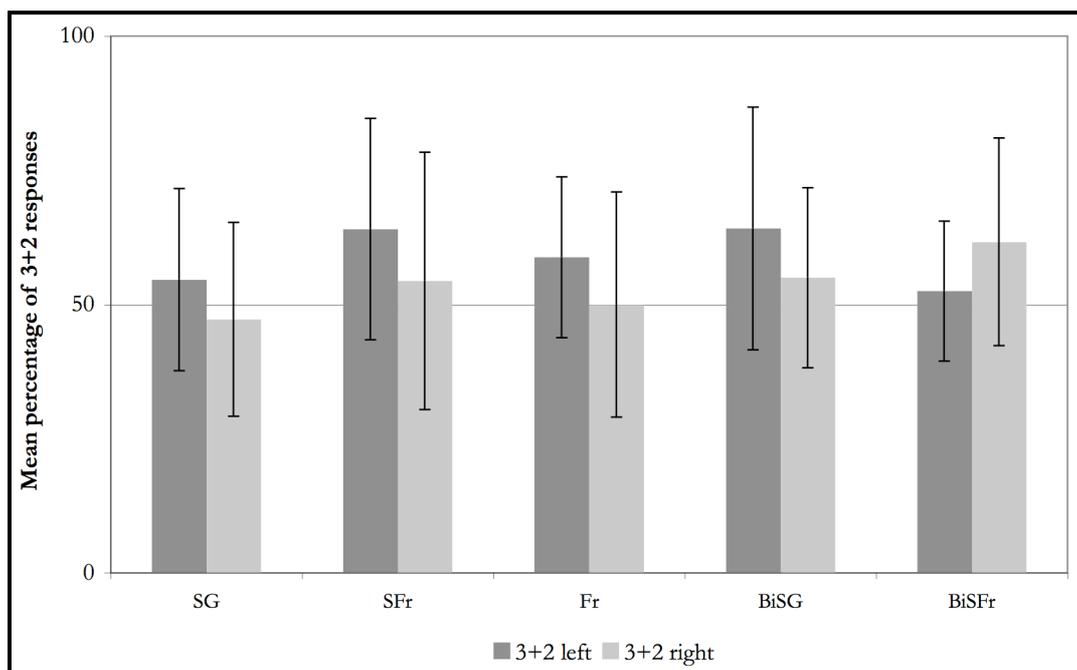


Figure 4-5 – Mean percentage (across subjects) of 3+2 responses to control stimuli (error bars: ± 1 standard deviation); 3+2 left/right refers to which side of the screen subjects saw the 3+2 grouping

If no bias occurred, responses would be 50% ‘3+2’ (and 50% ‘2+3’). We see a preference for 3+2 grouping in most of the ten subject groups, but in only four groups was this bias significant according to binomial probabilities approximated from the standard normal distribution (see Table 4-9; the BiSG 3+2 left group are the same subjects as the BiSFr 3+2 right group).

	SG		SFr		Fr		BiSG		BiSFr	
	3+2		3+2		3+2		3+2		3+2	
	<i>left</i>	<i>right</i>								
3+2 responses	118	102	146	124	127	108	77	66	63	74
Number of control trials	216	216	228	228	216	216	120	120	120	120
\bar{z}	1.36	-0.82	4.24	1.32	2.59	0.00	3.10	1.10	0.55	2.56
sig.	NS	NS	***	NS	*	NS	**	NS	NS	*
\bar{z} z-score for the binomial probability approximated from the standard normal distribution significance: *** $p < 0.0001$; ** $p < 0.001$; * $p < 0.01$; NS, $p > 0.05$										

Table 4-9 – Frequency of 3+2 responses out of total number of control trials per language group, and significance tests

House (1990) also found a 3+2 bias (67% and 77% 3+2 responses in two types of control stimuli). He suggested that this could have a cultural cause, since Swedish postal codes and telephone numbers are a 3+2 format, or it could be an experimental artefact, since the 3+2 grouping was nearer the right edge of the paper on which subjects marked the boundary, which was possibly easier (for right-handers). In the present experiment, which used a computer-based method that was visually counterbalanced across subjects, who had a different cultural background, the bias was greatly reduced (56% 3+2 responses over all subjects). Nevertheless, since bias occurred in some groups, a ‘difference score’ calculation explained by House (1990: 94) was also adopted here:

‘A difference score representing the percent 3+2 responses for stimulus ‘a’ minus the percent 3+2 responses for stimulus ‘b’ is presented for each stimulus pair. The difference scores represent the relative contribution of each variable to the perception of the [rhythmic-group] boundary where a higher difference score for a stimulus pair means a greater contribution to the perception of phrasing of the variable manipulated in that pair. For example, if 100% of the responses for stimulus ‘a’ in a pair were 3+2 and 0% of the responses for stimulus ‘b’ in the same pair were 3+2 (i.e. 100% 2+3 responses) the difference score would be 100-0=100 giving us the maximum contribution of the variable which was manipulated. If, however, say 70% of the responses for stimulus ‘a’ were 3+2 and 50% for stimulus ‘b’ were 3+2 we would get a difference score of 70-50=20 which would mean that there was very little contribution from the manipulated variable.’

4.6.2 Initial analysis: all variables

The data-set was inspected visually using Figure 4-6. The area of individual rectangles represents the difference score (henceforth DS), summed over subjects, for each level of three variables (language(s), cue(s), pattern). Imagine a mosaic of completely equally sized rectangles: this would represent a data-set in which the DS was not affected by any of these variables. However, we see inequality between certain rectangles (e.g. ‘pitch1/2 & length conflicting’ are smaller than ‘pitch1/2 & length accordant’, more so for SGs and bilinguals than (S)Fr), so these variables influenced listeners’ responses.

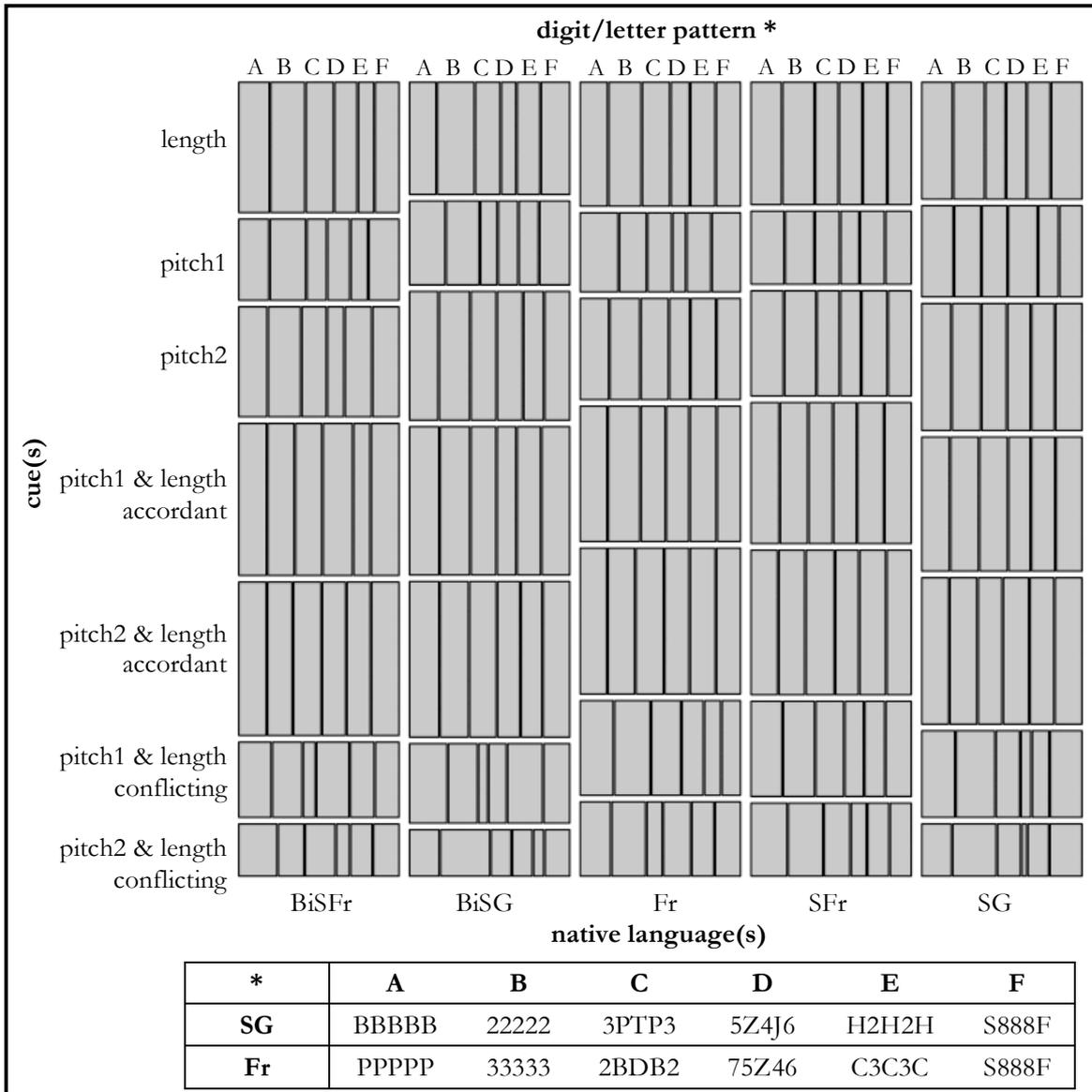


Figure 4-6 – Mosaic diagram: areas of rectangles represent the DS (summed over subjects) for each level of three variables (language(s), cue(s), pattern)¹

¹ Since language groups were not equally-sized, each rectangle is proportional to the total DS across all ‘cue(s)’ levels in the respective language group. Occasionally subjects responded 2+3 for stimulus ‘a’ (0)

To explore the effect of these variables statistically, a logistic regression analysis was conducted. Since each subject responded to several stimuli, a mixed model with both random and fixed effects must be fitted (Baayen 2008, Garson 2009). The random effect was *subject*, which introduced adjustments to the intercept grouped by each subject. The fixed effects were *native language(s)* (5 levels), *cue(s)* (7 levels), *digit/letter pattern* (6 levels), *order* (2 levels). The dependent variable was *response* ('3+2' = 1; '2+3' = 0); it was not necessary to use DS as the dependent here, since instead *order* accounted for any bias. The 'lmer' function within the R software environment was used (as in chapter 3)². Table 4-10 displays the output; the right-most column, which displays significance values for fixed effects, is of most interest. Each level of each variable was compared to the baseline level (rows without figures). We see that:

- SFr responses differed significantly from SG responses;
- responses to 'pitch1/2' and 'pitch1/2 & length conflicting' stimuli differed significantly from responses to 'length' stimuli;
- unsurprisingly, responses differed significantly depending on whether the second or third syllable was manipulated (order a/b);
- responses to three out of five digit/letter patterns differed significantly from responses to BBBBB/PPPPP stimuli.

and 3+2 for stimulus 'b' (1), hence a negative DS: $0-1=-1$. These were treated as 0 in this diagram, as otherwise some 'pitch1/2 & length conflicting' rectangles would have needed 'negative' areas. This was nevertheless a useful approximation for a preliminary inspection of the data (analysis of responses to 'conflicting' stimuli comes in §4.6.4.3).

² The formula was:

```
> expt2.lmer = lmer(response_3+2 ~ language + cue + pattern + order + (1|subject), data =
expt2, family = "binomial")
```

	<i>Fixed effects</i>	Estimate	Std. Error	ζ	p
	Intercept	1.206	0.126	9.55	<0.0001***
native language(s)	SG				
	SFr	0.288	0.138	2.09	0.037*
	Fr	-0.038	0.140	-0.27	0.787
	BiSG	0.037	0.167	0.22	0.824
	BiSFr	0.087	0.167	0.52	0.604
cue(s)	length				
	pitch1	0.520	0.084	6.19	<0.0001***
	pitch2	0.295	0.084	3.52	<0.001***
	pitch1 & length accordant	0.049	0.084	0.59	0.558
	pitch2 & length accordant	0.056	0.084	0.67	0.503
	pitch1 & length conflicting	-0.161	0.084	-1.93	0.050*
	pitch2 & length conflicting	-0.256	0.084	-3.05	0.002**
order	order a				
	order b	-2.796	0.047	-60.07	<0.0001***
digit/letter pattern	BBBBB, PPPPP				
	22222, 33333	0.081	0.078	1.05	0.296
	3PTP3, 2BDB2	0.410	0.078	5.27	<0.0001***
	5Z4J6, 75Z46	0.162	0.078	2.09	0.036*
	H2H2H, C3C3C	-0.199	0.078	-2.56	0.011*
	S888F, S888F	-0.084	0.078	-1.08	0.278
significance: *** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$					

Table 4-10 – Output of regression model

Now that the significant effects of variables have been highlighted, the following sections report a further analysis of the ‘digit/letter pattern’ variable, and then the main analysis, which concerns ‘native language(s)’ and ‘cue(s)’. Henceforth, the dependent variable is DS, to account for any 3+2 bias.

4.6.3 Digit/letter pattern

For each digit/letter pattern, subjects’ DSs were averaged across all cue conditions (see appendix 8.2.2), and then input to a two-way mixed-measures ANOVA with the factors *language*

(SG, SFr, Fr: between groups³) and *pattern* (x6: repeated-measures). No main effect of *language* occurred [$F(2,107)=1.025$, $p>0.05$], nor an interaction of *pattern* \times *language* [$F(9.189,491.621)=1.201$, $p>0.05$], so the SG and Fr stimuli, even though they were segmentally different, had a cross-linguistically equivalent effect on responses. There was a main effect of *pattern* [$F(4.595,491.621)=38.532$, $p<0.0001$], and to explore this further, post-hoc pairwise comparisons were computed. With a Bonferroni-adjusted alpha level [$p<0.003$ (0.05/15)], a significant difference occurred between the two digit/letter patterns for the cells marked * in Table 4-11. Mean DSs were significantly higher for ‘identical’ than ‘varied’ stimuli, and not significantly different between the two ‘identical’ types. Stimuli with five segmentally different syllables (5Z4J6/75Z46) had the lowest mean DS, which differed significantly from two of the three other ‘varied’ types.

	BBBBB/ PPPPP	22222/ 33333	3PTP3/ 2BDB2	5Z4J6/ 75Z46	H2H2H/ C3C3C	S888F/ S888F
BBBBB/ PPPPP		non- sig.	*	*	*	*
22222/ 33333			*	*	*	*
3PTP3/ 2BDB2				*	non-sig.	non-sig.
5Z4J6/ 75Z46					non-sig.	*
H2H2H/ C3C3C						non-sig.
S888F/ S888F						

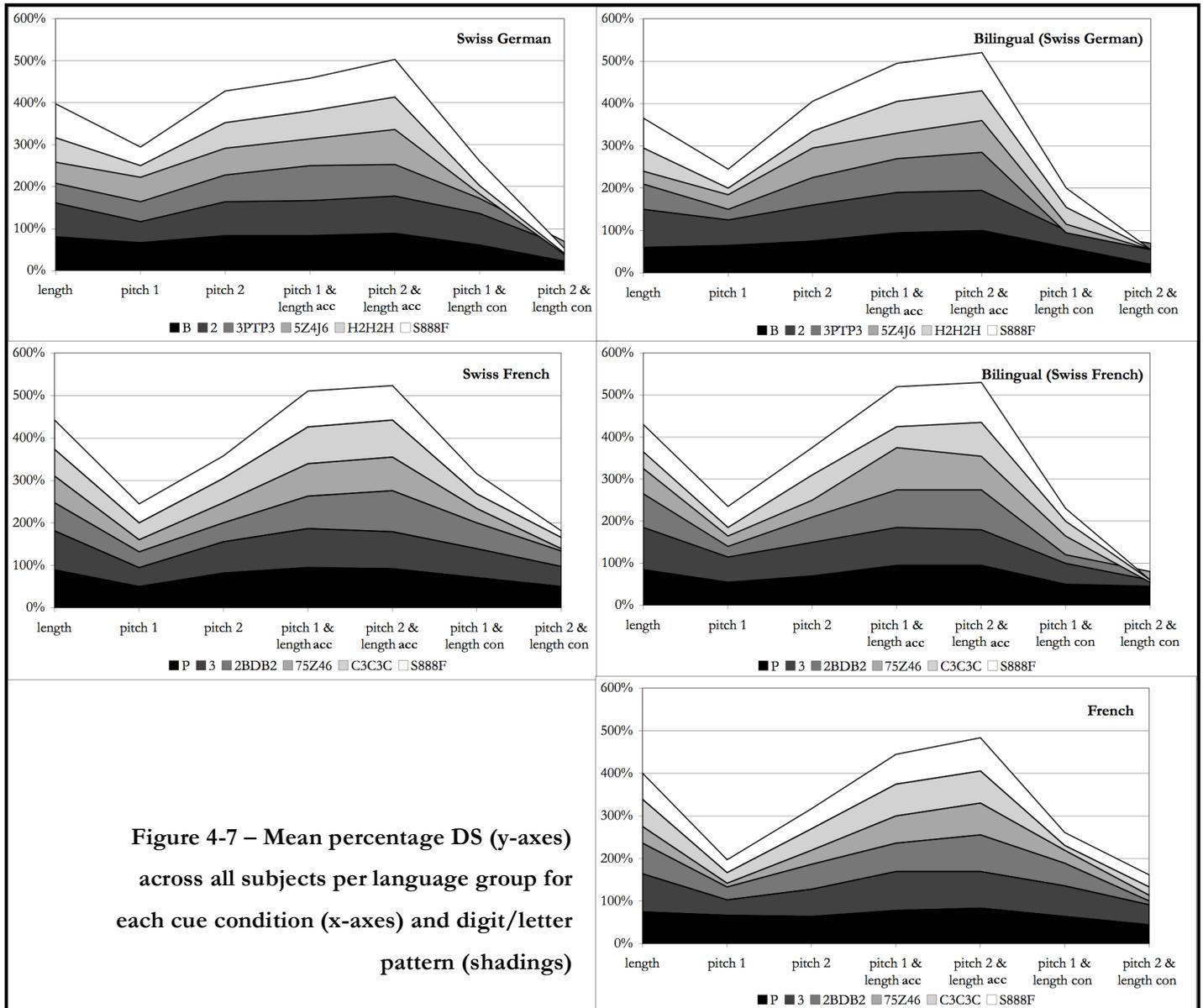
Table 4-11 – Post-hoc pairwise comparisons for *pattern*

A repeated-measures ANOVA on the bilinguals’ averaged DSs was computed with the factors *language* (BiSG, BiSFr) and *pattern* (x6). As for monolinguals, no main effect of *language* occurred [$F(1,19)=0.8$, $p>0.05$], nor an interaction of *language* \times *pattern* [$F(5,95)=1.215$, $p>0.05$], and a main effect of *pattern* occurred [$F(5,95)=6.55$, $p<0.0001$]. Post-hoc pairwise comparisons for *pattern* with Bonferroni-adjusted alpha levels revealed an overall picture similar to the monolinguals. The only noteworthy difference was a lack of significant difference between S888F and both ‘identical’ types.

³ Initially just the monolinguals’ responses were analysed, since the bilinguals formed a repeated-measures design.

4.6.4 Main analysis: cue(s); native language(s)

Figure 4-7 shows the mean DSs (across all subjects per language group). A 100% DS indicates a maximum effect of that variable in subjects' rhythmic-group judgments, whereas a 0% DS indicates no effect. Each cue condition adds up to maximally 600% (100% for each digit/letter pattern)⁴.



⁴ On the SG and both bilingual graphs, there is some overlap of the lines representing DSs for different digit/letter patterns in the 'conflicting' conditions. Where a line which top-borders a lighter shading falls below a line which top-borders a darker shading, this indicates a negative DS for the digit/letter pattern with lighter shading. For 'conflicting' stimuli, a negative mean DS means that pitch was used as a cue (i.e. the f_0 rise came at the end of a subject's perceived RG) more often than length was used as a cue, whereas a positive mean DS means that length was used as a cue more often than pitch was (compare Figure 4-4 and DS explanation in §4.6.1). §4.6.4.3 further discusses responses to 'conflicting' stimuli.

The ‘landscape’ is similar for all five graphs. A peak on the left for ‘length’, followed by a valley for ‘pitch1’, then a rise to the ‘accordant’ conditions, and a fall to the ‘conflicting’ conditions. Some subtle differences appear between language groups. For monolinguals, the ‘pitch1’ valley is deepest for Fr, slightly less for SFr and much shallower for SG, i.e. ‘pitch1’ influenced responses more in SG than in (S)Fr. The two ‘accordant’ conditions form more of a plateau for (S)Fr compared to a peak for SG, which (on the graph) results from the fact that ‘pitch2’ influenced responses more than ‘length’ in SG. The three left-most conditions are ranked (most to least effect on responses) as ‘pitch2’ > ‘length’ > ‘pitch1’ in SG, and ‘length’ > ‘pitch2’ > ‘pitch1’ in (S)Fr. The ‘conflicting’ conditions generally influenced responses more for (S)Fr than SG. The bilinguals behaved similarly to monolinguals when listening to the respective stimuli, as the BiSG and BiSFr graphs are more similar to the SG and (S)Fr graphs respectively. In the following analyses of the interaction between the ‘cue(s)’ and ‘native language(s)’ variables (which was the experiment’s aim), it was necessary to average subjects’ DSs across the digit/letter pattern conditions. This was justified because the same digit/letter patterns occurred in all cue conditions.

4.6.4.1 Monolinguals

A two-way mixed-measures ANOVA was conducted with the factors *language* (SG, SFr, Fr: between groups) and *cue(s)* (x7: repeated-measures). No main effect of *language* occurred [$F(2,107)=1.021$, $p>0.05$], but there was a significant interaction of *cue(s)* \times *language* [$F(7.330,392.138)=2.222$, $p=0.029$], and a main effect of *cue(s)* [$F(3.665,392.138)=72.310$, $p<0.0001$]. For the *cue(s)* \times *language* interaction, individual contrasts comparing ‘length’ to every other condition revealed that the only significant difference was between ‘length’ and ‘pitch2’ (Table 4-12). This shows that the cross-linguistic differential ranking of pitch and length observed in Figure 4-7 is significant: for SG, ‘pitch2’ > ‘length’ (> ‘pitch1’); for (S)Fr, ‘length’ > ‘pitch2’ (> ‘pitch1’).

	Source	df	Mean Square	<i>F</i>	<i>p</i>
cue(s) \times language	length vs. pitch1	2	0.319	2.466	0.090
	length vs. pitch2	2	0.441	4.643	0.012*
	length vs. pitch1 & length acc	2	0.015	0.157	0.855
	length vs. pitch2 & length acc	2	0.018	0.286	0.752
	length vs. pitch1 & length con	2	0.004	0.038	0.963
	length vs. pitch2 & length con	2	0.292	1.578	0.211
significance: * $p<0.05$					

Table 4-12 – Planned contrasts for the *cue(s)* \times *language* interaction

To verify that no between-variety difference occurred between SFr and Fr, another two-way mixed-measures ANOVA was calculated for *language* and *cue(s)* with only SFr and Fr. No

main effect of *language* occurred [$F(1,72)=2.722, p>0.05$], nor an interaction of *language* \times *cue(s)* [$F(3.808,274.173)=0.132, p>0.05$].

To explore the main effect of *cue(s)* further, post-hoc pairwise comparisons of all conditions were computed. With a Bonferroni-adjusted alpha level [$p<0.002 (0.05/21)$], a significant difference occurred between the two cue conditions for the cells marked * in Table 4-13. For almost every cue condition, the mean DS was significantly higher/lower than for all other conditions. The cue or cues which occurred on the second and/or third syllable of a five-syllable sequence profoundly influenced listeners' perception of rhythmic groups. The mean DS was not significantly different between 'length' and 'pitch2', which was the only comparison to reach significance in the *cue(s)* \times *language* interaction reported above. Therefore, only when native language is taken into account did increased duration and substantially rising f0 differentially influence responses.

	length	pitch1	pitch2	pitch1 & length acc	pitch2 & length acc	pitch1 & length con	pitch2 & length con
length		*	non-sig.	*	*	*	*
pitch1			*	*	*	non-sig.	*
pitch2				*	*	*	*
pitch1 & length acc					non-sig.	*	*
pitch2 & length acc						*	*
pitch1 & length con							*
pitch2 & length con							

Table 4-13 – Post-hoc pairwise comparisons for *cue(s)*

4.6.4.2 Bilinguals

To further explore the native language(s) variable, bilinguals' responses were analysed. A two-way repeated-measures ANOVA was conducted with the factors *language* (BiSG, BiSFr) and *cue(s)* (x7). There was a main effect of *cue(s)* [$F(2.248,42.713)=35.569, p<0.0001$], but not *language* [$F(1,19)=0.8, p>0.05$], as for monolinguals. The *cue(s)* \times *language* interaction was not significant overall [$F(6,114)=0.742, p>0.05$], but individual contrasts for this interaction, comparing 'length' to every other condition, revealed a significant difference between 'length' and 'pitch2' [$F(1,19)=6.241, p<0.05$], as for monolinguals. When bilinguals listened to SG five-syllable sequences, a substantial f0 rise in the second/third syllable influenced their rhythmic-group perception more than an increased duration did, whereas when they listened to Fr, an increased duration influenced rhythmic-group perception more than a substantial f0 rise did.

The main effect of *cue(s)* was explored further with post-hoc pairwise comparisons of all conditions. With a Bonferroni-adjusted alpha level [$p < 0.001$ ($0.05/42$)], there was a significant difference between the two cue conditions for the cells marked * in Table 4-14. Fewer significant differences occurred than for the monolinguals (compare Table 4-13). Marked in bold are cases in which a significant difference occurred in one language but not the other.

	length	pitch1		pitch2		pitch1 & length acc		pitch2 & length acc		pitch1 & length con		pitch2 & length con	
<i>Language (of stimuli)</i>		<i>SG</i>	<i>Fr</i>	<i>SG</i>	<i>Fr</i>	<i>SG</i>	<i>Fr</i>	<i>SG</i>	<i>Fr</i>	<i>SG</i>	<i>Fr</i>	<i>SG</i>	<i>Fr</i>
length		non-sig.	*	non-sig.	non-sig.	*	non-sig.	*	non-sig.	*	*	*	*
pitch1				*	non-sig.	*	*	*	*	non-sig.	non-sig.	non-sig.	non-sig.
pitch2						non-sig.	*	non-sig.	*	non-sig.	non-sig.	*	*
pitch1 & length acc								non-sig.	non-sig.	*	*	*	*
pitch2 & length acc										*	*	*	*
pitch1 & length con												non-sig.	non-sig.
pitch2 & length con													

Table 4-14 – Post-hoc pairwise comparisons for *cue(s)* × *language* (bilinguals)

Consider the bold cases: ‘pitch1’ had a significantly lower mean DS than ‘length’ for Fr but not SG, whereas ‘pitch2’ had a significantly higher mean DS than ‘pitch1’ for SG but not Fr; ‘length’ had a significantly lower mean DS than the two ‘accordant’ conditions for SG but not Fr, whereas ‘pitch2’ had a significantly lower mean DS than the two ‘accordant’ conditions for Fr but not SG. These significant differences all accord with the finding that when bilinguals heard SG, a substantial f0 rise influenced rhythmic-group perception more than an increased duration did, and the opposite when they heard Fr.

Then a two-way mixed-measures ANOVA was conducted with the factors *group* (SG, BiSG: between groups) and *cue(s)* (x7: repeated-measures). Likewise a two-way mixed-measures ANOVA was conducted with the factors *group* (SFr, Fr, BiSFr: between groups) and *cue(s)* (x7: repeated-measures) (Table 4-15). For both ANOVAs, there was a main effect of *cue(s)* (as above), but not *group*, nor an interaction of *cue(s)* × *group*, so there was little difference in responses between the bilinguals and each respective monolingual group.

	Source	df	Mean Square	F	p
SG groups	cue(s)	3.018	7.253	56.718	<0.0001*
	cue(s) × group	3.018	0.090	0.701	0.553
	Error	162.959	0.128		
	group	1	0.009	0.154	0.696
	Error	54	0.060		
(S)Fr groups	cue(s)	3.715	7.772	64.105	<0.0001*
	cue(s) × group	7.430	0.124	1.025	0.415
	Error	338.083	0.121		
	group	2	0.052	1.496	0.229
	Error	91	0.035		
<ul style="list-style-type: none"> • A Greenhouse-Geisser correction was applied since sphericity could not be assumed (Mauchly's test, $p < 0.001$) • significance: * $p < 0.0001$ 					

Table 4-15 – ANOVA output: *cue(s) × group* (comparing monolinguals and bilinguals; one ANOVA for each language)

4.6.4.3 ‘Conflicting’ stimuli

The significantly lower DSs for ‘conflicting’ stimuli show that listeners were confused by what they heard. From responses to these stimuli, we can see which cue ‘won’ when listeners were faced with choosing between length and pitch cues. Figure 4-8 shows how often subjects used either length or pitch as a cue, i.e. increased duration or rising f_0 (respectively) was group-final in their perceived first rhythmic group.

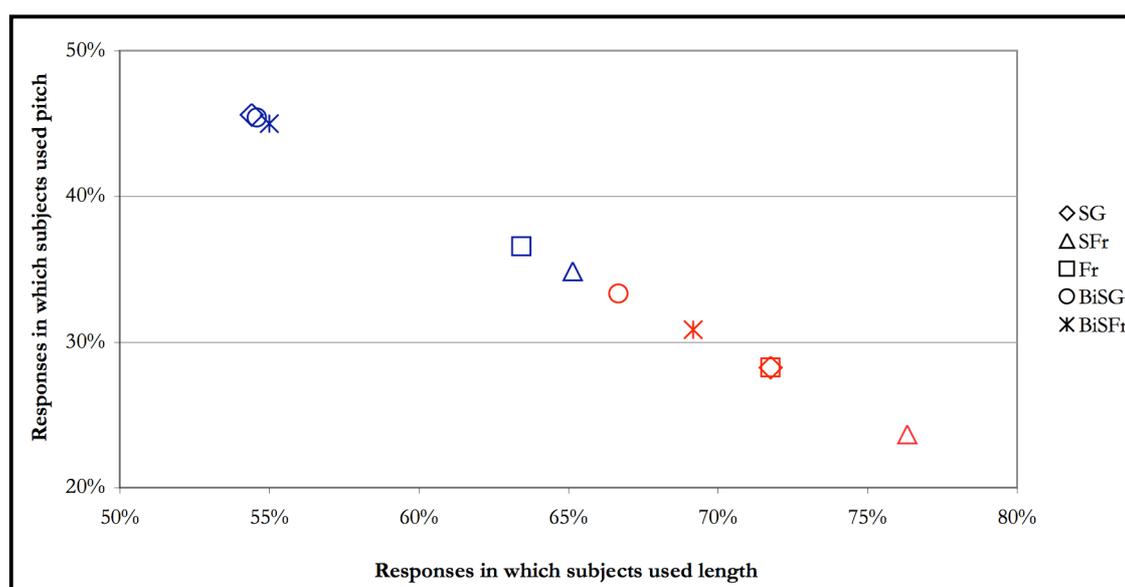


Figure 4-8 – Percentage of responses to ‘conflicting’ stimuli in which subjects used either length or pitch; red, ‘pitch1 & length conflicting’, blue, ‘pitch2 & length conflicting’

According to binomial probabilities approximated from the standard normal distribution, subjects used length as a cue significantly more often (i.e. pitch significantly less often) than chance (50%) in all groups [$z > 5$, $p < 0.001$], except SG, BiSG and BiSFr subjects in the ‘pitch2 & length conflicting’ condition (top-left cluster). Generally, pitch was used as a cue more often for stimuli with a higher f0 excursion (blue) than for stimuli with a lower f0 excursion (red). Figure 4-9 shows that it was not the case that some individuals only ever used length and some only pitch; most subjects sometimes used length and sometimes pitch, each cue in around one to two thirds of stimuli.

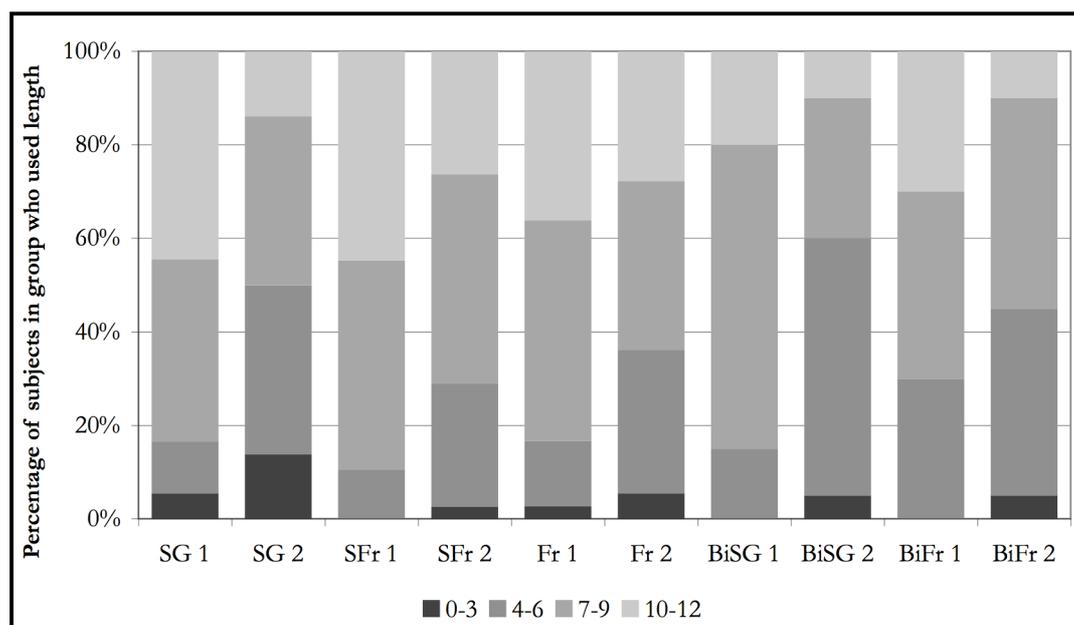


Figure 4-9 – Percentage of subjects who used length in 0-3, 4-6, 7-9 and 10-12 of the 12 ‘conflicting’ stimuli for each pitch condition (1 or 2)

4.7 Discussion

It was predicted that listeners would locate rhythmic-group boundaries using length and pitch cues, so that: when only one cue was available, the mean DS would be relatively high; when both cues were available on one syllable, the mean DS would be even higher; when the cues were available on conflicting syllables around the boundary location, the mean DS would be much lower; when neither cue was available, responses would be around chance level. This predicted pattern of results occurred for all language groups. The following discussion first interprets the evidence for the perceptual interdependence of duration and f0, and then addresses cross-linguistic variation and other interesting findings.

4.7.1 Interdependence of duration and f0

For monolinguals, when both increased duration and dynamic f0 occurred simultaneously (‘accordant’ stimuli), listeners were significantly more convinced of the rhythmic-group boundary location (i.e. higher DS) than when only one cue was available (‘length’ and

'pitch2' stimuli). When the cues conflicted on adjacent syllables ('conflicting' stimuli), listeners were significantly less convinced of the boundary location (i.e. lower DS) than when only one cue was available, and did not attend to the same cue in every 'conflicting' stimulus. Taken together, these findings demonstrate that increased duration and dynamic f_0 are interdependent cues to rhythmic groups. Here we approach the subject of phonetic trading relations, which are usually discussed in terms of cues to segmental contrasts, e.g. plosives. Repp (1982: 87) stated:

[...] virtually every phonetic contrast is cued by several distinct acoustic properties of the speech signal. It follows that, within limits set by the relative perceptual weights and by the ranges of effectiveness of these cues, a change in the setting of one cue (which, by itself, would have led to a change in the phonetic percept) can be offset by an opposed change in the setting of another cue so as to maintain the original phonetic percept. This is a phonetic trading relation. [...] neither cue is perceived in isolation; rather, they are perceived together and integrated into a unitary phonetic percept.'

In this experiment, the two cues are increased duration and rising f_0 , and the contrast is not segmental but rather the categorical location of a rhythmic-group boundary after the second or third syllable in a five-syllable sequence. When one syllable has increased length and (sufficiently) rising pitch, the cues are perceived together as being accordant, and integrated into the phonetic percept of a boundary after that syllable. When one syllable with increased length precedes another with (sufficiently) rising pitch, or vice versa, the cues are perceived together, but as being conflicting; the phonetic percept resulting from integration of the cues is less clearly in one category or the other. As a rhythmic-group cue, increased length may have a greater perceptual weight than rising pitch for (S)Frs, but the opposite for SGs. Therefore, the percept of a boundary may depend more on the duration setting than the f_0 setting (present too) for (S)Frs, and the opposite for SGs. We could also interpret in trading relations terms the results for stimuli with only one cue present. Increased length may be the most usual rhythmic-group cue, but when this is missing, rising pitch (if sufficiently large in excursion) can be equivalent, and so maintains the percept of a boundary following that syllable. Depending on native language, the relationship may be reversed, i.e. rising pitch is the most usual cue. Experiment 1's finding (chapter 3) adds to this possibility; listeners perceived syllables with dynamic f_0 as longer than those with level f_0 , so the physical correlate of perceived length is sometimes increased physical duration and sometimes dynamic f_0 . For a more definitive conclusion, another experiment could be conducted whereby subjects indicate which was the longest syllable in five-syllable sequences with tonal and durational manipulations. Subjects in this experiment were not asked which syllable was longest, because it would have distracted their attention from the grouping task.

4.7.2 Native language

In cross-linguistic experiments, it is challenging to create stimuli that are comparable between languages and equally appropriate for all listeners (Beddor and Gottfried 1995). Here we can assume that the SG and Fr stimuli, though segmentally different, were cross-linguistically equivalent in terms of the responses they elicited, because no significant interaction of native language and digit/letter pattern occurred. No prediction was made concerning native language, because it was unclear whether the use of length and pitch cues would differ between language groups, since native language did not affect the interdependence of f_0 and duration in experiment 1. The present results now prompt discussion.

Increased duration and dynamic f_0 (with substantial excursion) can each effectively cue rhythmic-group boundaries when the other is absent, but whether one is more effective than the other depends on native language. There was a significant interaction of language and cue condition for monolinguals, and some notable differences occurred in the effect of length and pitch cues between the two languages for bilinguals. Rising pitch (with sufficiently large excursion) was a significantly more important cue than increased length for monolingual SGs, whereas increased length was a significantly more important cue than rising pitch for monolingual (S)Fr. SGs were more sensitive than (S)Fr. even to small (10Hz) rises. Responses to ‘conflicting’ stimuli (Figure 4-8) showed that (S)Fr. generally used length as a cue more, and pitch less, than SGs. Only when native language was taken into account was there a significant difference between length and pitch cues in terms of their effect on responses. Bilinguals also showed this hierarchy of ‘pitch > length’ for SG stimuli and ‘length > pitch’ for Fr stimuli.

The finding that tonal and durational cues are more important for SGs and (S)Fr. respectively can be interpreted with reference to the prosodic characteristics of each language (details and references in chapter 2). In Fr, each rhythmic group always has a final prominent syllable, which has dynamic f_0 and is lengthened; rhythmic groups also have an optional initial prominent syllable, with an f_0 rise but no lengthening. According to Di Cristo (2000: 40), the general consensus is that group-initial prominence is probably cued by dynamic f_0 , whereas the primary cue to group-final prominence is probably syllable lengthening. Thus it is logical that (S)Fr. perceive increased duration as a stronger (right-hand) rhythmic-group-boundary cue than rising f_0 . In SG, syllables with lexical stress, which are mostly content-word-initial, have a rising f_0 that often starts late and continues into the following (perceptually non-prominent) syllable, so this rise may often end towards the right-hand boundary of rhythmic groups. Phrase-final syllables are lengthened, as are initial and prominent (medial) ones to some extent. SGs may perceive rising pitch as a stronger rhythmic-group boundary cue, because increased duration may be a less specific cue to group-finality than rising f_0 . This explanation must remain tentative because perceptual data on SG and (S)Fr. prominence cues is limited. These results nevertheless provide insight on boundary cues in these languages.

There is little evidence here for a difference between SFr and Fr. In Figure 4-7, these varieties' slightly different 'landscapes' result mainly from 'pitch1' and 'pitch1 & length accordant' influencing responses (non-significantly) more in SFr than Fr, i.e. SFr seemed more sensitive to smaller pitch excursions than Frs. However, *all* cue conditions generally influenced responses more in SFr than Fr. When duration and f0 conflicted as cues on adjacent syllables, SFr used length as a cue more, and pitch less, than Frs. Perhaps this is related to the fact that dynamic f0 may occur more towards the centre of rhythmic groups in SFr rather than at the boundaries, as Miller (2007) found that group-final rises started earlier, and group-initial rises started later, in SFr than Fr.

In terms of which cue more effectively signalled rhythmic-group boundaries, bilinguals responded like SGs when listening to SG stimuli and like (S)Fr when listening to Fr stimuli. Given that prosody perception develops in early infancy, the responses of these bilinguals, who acquired both languages early in life, provide more evidence that native language influences the relative weighting of duration and f0 in the interdependence of these rhythmic-group cues. Some interesting differences between bilinguals and monolinguals occurred in responses to 'conflicting' stimuli: when f0 excursion was small (10Hz) bilinguals used as a cue (for both languages) length less and pitch more often than monolinguals did; when f0 excursion was larger (30Hz), bilinguals used as a cue length and pitch a similar number of times (for both languages), like the SGs (Figure 4-8); more bilingual than monolingual individuals used length and pitch as a cue in fairly equal proportions across the 'conflicting' stimuli (Figure 4-9). Generally bilinguals had a means of separating for each language their strategy of rhythmic-group perception, so as to respond like monolinguals. Yet when the task got difficult due to conflicting cues, listeners' cognitive load was increased. In bilinguals this may have led to a decreased capacity to maintain separation between the languages and to associate responses with one specific prosody, hence their relatively equal use of length and pitch for the stimuli in which the cues conflicted.

4.7.3 Other findings

Subjects were more convinced of the boundary location in stimuli with greater f0 excursion on the relevant syllable. 'Pitch1' (smaller excursion) had a significantly lower mean DS than 'pitch2' for monolinguals overall, and for bilinguals when listening to SG but not Fr. For 'accordant' stimuli, those with larger excursion generally resulted in (non-significantly) greater DSs than those with smaller excursion. For 'conflicting' stimuli, those with smaller excursion generally resulted in (non-significantly) greater DSs than those with larger excursion, i.e. subjects found the smaller f0 change less confusing. The larger 30Hz excursion was based on the recordings' values (see §4.5.2.3); the smaller 10Hz excursion was chosen because previous experiments have shown that it would unlikely be perceived as a rise. Rossi (1971) found that the

threshold for perceiving a rising f_0 was 19Hz for 200ms long vowels, which he compared with Sergeant and Harris' (1962) finding of 16Hz⁵. In the present experiment, the smaller f_0 rises were less perceptible, and (perhaps depending on native language) were considered as micro-fluctuations, so generally did not signal linguistic (i.e. rhythmic-group) structure, unlike larger rises.

The effects of increased length and dynamic pitch as rhythmic-group cues were significantly greater in series of identical syllables than series with segmentally varied syllables; for varied sequences, it made little difference to responses whether the five syllables alternated between two/three segmental structures (e.g. 2BDB2) or all had different segments (e.g. 5Z4J6). Generally, perceiving series of varied syllables should present a greater cognitive load for listeners than identical syllables would, because the various vowels and consonants to process differ in spectral properties, including intrinsic pitch which leads to f_0 micro-perturbations in the continuous signal. Therefore, less cognitive capacity remains to use prosodic cues (e.g. increased duration and dynamic f_0) as effectively when processing 'varied' stimuli rather than 'identical' stimuli. This accords with House (1990), who proposed that f_0 movement is optimally perceived in periods of spectral stability, when perception is not constrained by attention to spectral changes, and this applies at the syllable level and in marking phrase boundaries. The present experiment could extend this conclusion from tonal research to prosody in general; f_0 movement, increased duration (and conceivably other prosodic cues) could be optimally perceived depending on where they occur relative to spectral change. Segmentally varied stimuli were included to ensure that listeners heard a constantly changing signal, which is true of real speech, so the results are more generalisable to real speech perception than results from stimuli with identical syllables.

4.8 Conclusion

It is not always clear how we should define prosodic groups in speech production data, and perceptual data on this is lacking. This experiment has provided further evidence that, in continuous speech, groups of syllables defined by prosodic properties (f_0 , duration) are psychologically real. Subjects easily indicated that they heard groups when duration and f_0 manipulations were present, but their responses were at random when these cues were absent.

Experiment 1 (chapter 3) found a perceptual interdependence of f_0 and duration in stimuli isolated from linguistic context. This second experiment investigated whether f_0 and duration are interdependent rhythmic-group cues when listeners are given a linguistic context, and if so, whether this depends on native language. We saw that increased duration and dynamic

⁵ However, the just noticeable difference in pitch perception when comparing two tonally level vowels may be 6–12Hz or as low as 0.5Hz, according to Lehiste's (1970: 64) report of other experiments.

f0 can each cue rhythmic-group boundaries. Strong evidence that these cues are interdependent in rhythmic-group perception comes from the finding that two cues are significantly better than one when heard simultaneously, but significantly worse than one when heard in conflicting positions around the boundary location. Considered with experiment 1's results, increased duration and a (substantial) f0 rise may be perceptually equivalent in trading relations terms. When one cue is absent, the other cues the boundary, because the underlying percept correlates with two physical properties (increased duration and rising f0). Unlike in experiment 1, native language influenced the relative weighting of increased duration and dynamic f0 in these cues' interdependence. We can interpret the results as evidence that listeners' responses related to language-specific prosody. Furthermore, f0 excursion, to be an effective cue, must be great enough that it could be related to linguistic structure rather than an acoustic microfluctuation.

So far in this thesis, the purpose of the two experiments was to investigate whether f0 and duration are perceptually interdependent, because if so, this has implications for rhythm research which is duration-based. It is now clear that rhythm research should investigate f0 as well as duration, as these cues are interdependent. Speech rhythm is often defined along the lines of a regular pattern of alternating prominent and non-prominent units (e.g. Crystal 1985, Trask 1996). The grouping of syllables in continuous speech is relevant to rhythm, because groups build a structure in which the alternating pattern can occur. (Early psychologists recognised that how listeners group non-speech stimuli was important for rhythm research; see Fraisse 1982.) The experiment in this chapter essentially demonstrated a group-boundary phenomenon which is related to prominence, as listeners identified group boundaries by perceiving one syllable as more prominent than others. Prominence is not limited to boundaries, though this varies across languages; prominence and boundaries are highly correlated in Fr, but less so in SG. The experiment reported in the following chapter directly addresses the phenomenon of rhythm, by asking listeners to judge the rhythmicity of sentences in which f0 and duration are systematically manipulated to test whether these cues are interdependent in perceived rhythm. The stimuli were naturally produced sentences, which compared to the previous two experiments' stimuli were less prosodically stylised and linguistically more complex, so more like what is heard in everyday speech. This was important given that the perceptual interdependence of f0 and duration has manifested itself here more clearly with segmentally identical stimuli (unlike natural speech) than segmentally varied stimuli.

The interdependence of f0 and duration as cues to the perceived rhythmicity of sentences

5.1 Summary

The previous experiments found that f0 and duration are interdependent in the perception of isolated syllables and rhythmic groups. The experiment reported here investigates whether f0 and duration are interdependent perceptual cues when listeners have to judge the rhythmicity of longer utterances, and if so, whether this depends on native language. The stimuli are sentences, which we can expect listeners to process as they would any syntactically plausible utterance; duration and f0 are manipulated to test whether a deviant duration results in a less natural-sounding rhythm than a deviant f0 movement, or vice versa. The results demonstrate that duration and f0 are interdependent cues to perceived rhythm, and that the relative significance of a non-deviant duration and non-deviant f0 excursion depends on native language.

5.2 Previous research

5.2.1 How to test rhythm perception

Firstly, the question is whether naïve subjects are aware of speech rhythm, consciously or intuitively, and if so, whether they can judge sentence rhythmicity. Several rhythm perception experiments have given an affirmative answer to one or both of these questions. These experiments' methodologies are now discussed, categorised into four different types of task: tapping; interval adjustment; discrimination; and rating/judging rhythmicity.

Early experiments found that English subjects could tap their finger to a series of monosyllables (Miyake 1902) and to the prominent syllables in metrical verse (Wallin 1901). More recently, Allen (1972: 82, 92) found that English subjects could tap their finger quite precisely 'on the beat' of a specified syllable in spontaneous utterances, and place a click 'on the beat' of a specified syllable when they heard (several times) utterances with a superimposed adjustable click. Inter-subject variation was relatively high for tapping behaviour, and lower for click placement. Donovan and Darwin (1979) demonstrated that English subjects could tap their finger to four prominent syllables in a sentence. Scott et al. (1985), using English sentences prosodically identical to Donovan and Darwin's (1979) and prosodically equivalent French sentences, found that English and French subjects could tap their finger to four prominent syllables in sentences of their native language and the other language. French subjects' tapping was more temporally regular (i.e. periodic) overall and had significantly longer inter-tap intervals than the English subjects', which, Scott et al. (1985) suggested, might indicate that French subjects found the task less natural. Equally, English subjects (no French were tested) could tap to noise bursts, and to segmentally degraded (i.e. acoustically complex but linguistically meaningless) sentences. Tapping

was more temporally regular for speech (intelligible and unintelligible) than for noise bursts. From this Scott et al. (1985) argued that the more complex the acoustic signal, the harder the task, so the greater the bias towards evenly-spaced taps, and this demonstrated nothing about subjects' perception of rhythm in language.

Morton et al. (1976) showed that English listeners could adjust via a knob the timing of alternate monosyllables until they perceived each pair to occur at an equal interval. This experiment led to the proposal of P-centres in speech, and the task was later replicated in several P-centre experiments (e.g. Harsin 1997, Howell 1988, Marcus 1981, Pompino-Marschall 1989, Scott 1998). Donovan and Darwin (1979) found that English listeners could adjust via knobs the intervals between four noise bursts until they perceived the bursts as matching the timing of four prominent syllables in a synthesised sentence; the speech and noise bursts, which included intervening weaker bursts representing the unstressed syllables, were not heard simultaneously. English listeners could also perform the same task (though the results differed slightly from Donovan and Darwin's (1979) other experiment) when the following conditions were implemented: natural speech was used; all and only the prominent syllables began with [t], so listeners had to match 'the T's' rather than the 'syllable beats'; two intonation conditions were included (natural pitch contour and monotone); the weaker noise bursts were excluded; subjects were recorded repeating the sentence whilst adjusting the noise bursts.

In AAX discrimination tasks, Ramus et al. (2003) and White et al. (2007) told listeners that the stimuli they heard, which were segmentally degraded sentences from (real) rhythmically-differing languages, were exotic languages. Responses for whether an X stimulus was from the same/different language as the AA stimuli were barely above chance, from which White et al. (2007: 1011) concluded that 'this was a very difficult task for listeners'. The results of a classification task with prior training and similar stimuli led Ramus and Mehler (1999: 517) to conclude that 'the task demands sustained attention and an unusual effort to extract regularities'. Listeners' comments about their strategies were inconsistent and uninformative (Ramus and Mehler 1999: fn 4), which suggests that some found it difficult and/or behaved idiosyncratically. Such experiments, as well as being difficult for listeners, simply showed that listeners could (just about) distinguish rhythms in unnatural language (as had been found with prelingual infants, e.g. Mehler et al. 1988, Nazzi et al. 1998, Ramus et al. 2000), and revealed little about how acoustic properties contributed to listeners' impression of rhythmicity in language which conveys meaning (cf. Barry et al. 2009).

Benguerel and D'Arcy (1986) asked English, French and Japanese listeners to rate on a 7-point scale how 'speeded up' (1), 'regular' (4), 'slowed down' (7), or somewhere intermediate (2,3,5,6) they perceived stimuli to be which comprised six identical syllables or clicks with various interval timings. Inter- and intra-subject variation in responses was relatively high, so perhaps

some found the task harder than others, though between-language-group variation in responses was insignificant. Instead of timing, Grover and Terken (1995) asked Dutch listeners (some speech/sound experts and some naïve) to judge on a scale of 1 to 10 how ‘rhythmic’ eight- and nine-syllable nonsense utterances were, and to indicate the syllables they perceived as prominent and the location of group boundaries within the syllable series. These listeners could clearly distinguish degrees of rhythmicity (Grover and Terken 1995). Barry et al. (2009) presented English, German and Bulgarian listeners with an adjustable on-screen slide (rather than a discrete scale) to indicate how much an eight-syllable nonsense utterance was ‘more strongly rhythmical’ than another. Sliding upwards or downwards meant greater rhythmicity of utterance one or two respectively; sliding further from midway meant greater difference in strength of rhythmicity. Listeners defined their own range, and slider response positions were normalised. There was relatively high inter-subject variation and a few individuals behaved idiosyncratically, which, Barry et al. (2009) argued, could have resulted from the complexity of rhythm, or that the methodology could allow subjects to adopt various strategies which may (or may not) reflect those used in real-life. Nevertheless, individual subjects were ‘basically systematic’ in judging rhythmicity (Barry et al. 2009: 89). In interdisciplinary musicology-linguistics research, Magne et al. (2004) found that French subjects could judge whether the rhythm of the final tri-syllabic word/arpeggio in sentences and equivalent musical sequences was appropriate or not. Event-related potential (ERP) results demonstrated that unnatural rhythm was cognitively more complex to process than natural rhythm, in both speech and music.

All these experiments provide useful evidence that untrained subjects can demonstrate awareness of speech rhythm. Tapping, interval-adjustment and discrimination tasks may not help to answer the question of what it is in natural language that contributes to listeners’ impression of rhythm, whereas asking subjects to judge/rate rhythmicity may be a viable methodology for investigating this question. However, reports of intra- and inter-subject variation in rating judgements suggest that some subjects have clearer intuitions about rhythm in language than others, or that different task strategies may be possible (cf. Barry et al. 2009).

We should also consider that testing naïve (rather than trained) subjects has potential advantages. Miller (1984) asked phoneticians and non-phoneticians to indicate whether speech extracts from several languages were ‘syllable-timed’ or ‘stress-timed’. Although it is arguably impossible to explain these rhythm types to non-experts without biasing them, Miller (1984: 82) suggested that the non-phoneticians were less ‘influenced by received ideas’ than the phoneticians, since the non-phoneticians’ responses showed more inter-subject variation. This could also result from the phoneticians having more experience in general of listening to speech for experimental purposes. In discrimination tasks similar to those of Ramus et al. (2003) and White et al. (2007), van Dommelen (1987) found that responses were much higher above chance than in these other experiments (almost 100% for some stimuli). This performance difference

could have resulted from factors such as the instruction to discriminate ‘speakers’ (rather than ‘languages’ in Ramus et al. (2003) and White et al. (2007)), the method for segmentally degrading stimuli – laryngograph recordings (rather than ‘resynthesised’ speech in which all consonants were replaced with [s] and all vowels with [a]), or the fact that subjects were staff and students from a phonetics institute, who ‘coped surprisingly easily with the non-trivial task’ (van Dommelen 1987: 336). Phonetic training may increase listeners’ attentiveness to prominence cues, making an acoustic-based task easier for them, so their strategy in a rhythm-perception task may not be generalisable to listeners in a real-life situation (though not according to van Dommelen 1987). The very fact that untrained subjects are intuitively rather than consciously aware of speech rhythm makes their judgments appealing.

5.2.2 Evidence for interdependence of duration and f_0

The above experiments generally concerned certain trends in rhythm research (perceptual isochrony, rhythm typology) which focused on timing. Some hinted at the significance of intonation, by including monotonous stimuli and stimuli with their original f_0 contour (e.g. Donovan and Darwin 1979, Ramus et al. 2003).

An exception to this generalisation is Barry et al. (2009), who investigated the relative perceptual weight of durational, tonal, intensity and vowel quality properties in the perception of rhythm. Stimuli were eight-syllable nonsense utterances comprising four /**dɑ:də**/ feet. For each acoustic property, a trochaic (i.e. non-deviant) and a neutral (i.e. deviant) manipulation were defined: [longer + shorter] versus two durationally equal; [rising + level f_0] versus two tonally level (with declination); [louder + quieter] versus two equal in intensity; /**dɑ:də**/ versus /**dədə**/. Sixteen stimuli were created (every permutation of four properties being deviant or not) and presented in pairs. The first stimulus had all non-deviant properties (clearly trochaic rhythm) or all deviant properties (no rhythm cues), and the second differed from the first in one to three properties. Listeners indicated how much ‘more strongly rhythmical’ one stimulus was compared to the other. The results demonstrated an interaction of the four properties. Listeners tended to judge stimuli with two or three non-deviant properties as more strongly rhythmical than those with only one. For English and German listeners, non-deviant duration was most significant for a perceived rhythmical utterance, though non-deviant f_0 was also highly significant. For Bulgarians, non-deviant f_0 and duration were weighted almost equally. Grover and Terken’s (1995) experiment was similar. The stimuli comprised eight or nine [ma] syllables with various patterns of non-prominence and prominence, which involved syllable lengthening, a substantial f_0 rise, slightly increased amplitude and probably vowel quality. For their Dutch listeners, dynamic f_0 as well as increased duration (and possibly amplitude) contributed to perceived rhythmicality.

5.3 Pilot experiment

The present experiment was designed with the above experiments' methodologies in mind. Initially a pilot experiment was run which primarily tested whether listeners could judge rhythmicity with linguistically meaningful stimuli, since Barry et al. (2009) and Grover and Terken (1995) had demonstrated the viability of this task only with nonsense stimuli.

Equivalent stimuli were created in Fr and SG from five sentences recorded by a native speaker of each (these stimuli were not used in the main experiment). As Figure 5-1 illustrates, each sentence had nine syllables (represented by rectangles in the figure), three of which were prominent (shaded), though the position of the prominent syllables had to differ between languages (see §5.4.1). Each utterance's original f_0 was replaced by a stylised contour in *Praat* (Boersma and Weenik 2008-2009: version 5.1.01). A straight declination line was placed on the utterance, then stylised f_0 rises were added on prominent syllables, except the last one, which in the recordings always had a fall with a small excursion. The shape and positioning of the stylised rises reflected those observed in the recordings. From these stylised resynthesised versions of the utterances, eighty stimuli were created: (5 sentences x 8 duration manipulations) + (5 sentences x 8 f_0 manipulations). Duration or f_0 was manipulated on the second of the three prominent syllables. Duration ranged from 100ms to 450ms in 50ms steps, and f_0 excursion from 0Hz to 105Hz in 15Hz steps.

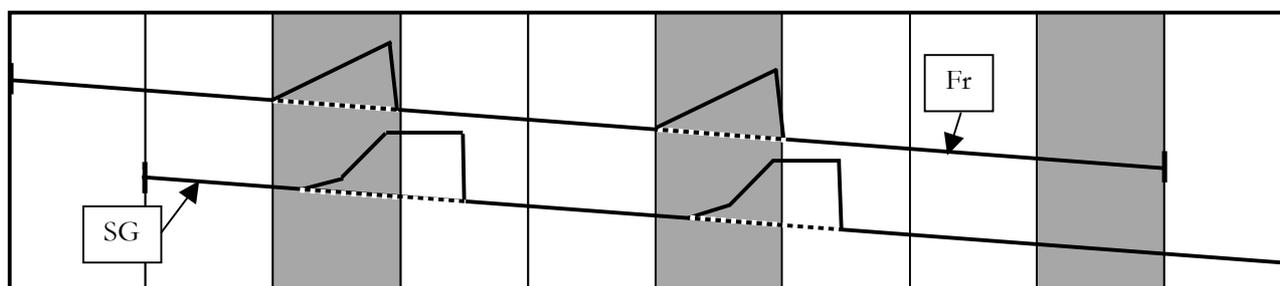


Figure 5-1 – f_0 contours for pilot stimuli

Six Fr and six SG subjects were tested with stimuli and instructions in their native language. Two tasks were trialled, a rating scale followed by an 'adjustment' task (with a break in between). These were run in *Praat*. A post-test questionnaire asked subjects to rate task difficulty. First, subjects heard the eighty (randomised) stimuli and indicated how 'rhythmic' each sounded on a scale of 1-9. There was a practice session, which presented examples from both extremes of duration/ f_0 manipulations, so subjects had reference points against which to rate rhythmicity. In the second task, each trial presented the eight versions (with durational or tonal manipulations) of one sentence. Subjects saw on screen a circle of eight boxes, each of which played a stimulus when clicked. As they clicked clockwise, the manipulated syllable increased in duration or f_0 excursion (starting from the top centre box). On hearing the sentence they thought was 'most rhythmic', they submitted this response. The circle idea originated in the

knob adjustment experiments detailed above: clicking clockwise, like turning a knob, increased one syllable's prominence (which was timing- or pitch-related) until subjects were satisfied.

The results primarily demonstrated that naïve listeners were willing and able to judge sentence rhythmality. The results also showed the values for duration and f₀ manipulations that listeners found acceptable, which proved useful in designing the main experiment's stimuli. In the post-test questionnaire, ratings of task difficulty (1 = very easy, 9 = very difficult) ranged from 1 to 8.5 (median = 5), so some subjects found it easier than others, but it was never impossible. All subjects except one wrote that the rating task was harder because they soon forgot the practice-session reference points after hearing sentences in the main session, whereas they thought the 'adjustment' task was easier because they had the control to compare several sentences together. Therefore, the 'adjustment' task was developed for the main experiment.

5.4 Main experiment: stimulus design

The hypothesis for this experiment is given after the stimulus design is explained.

5.4.1 Structure of sentences

As in experiment 2 (chapter 4), the SG and Fr stimuli were designed to be cross-linguistically comparable and equally appropriate for all listeners, though this was challenging (cf. Beddor and Gottfried 1995). Thirty sentences per language were constructed (appendix 8.3.1), then a native speaker checked them. In terms of prosodic structure, all sentences were identical within languages and, allowing for language-specific phonology, equivalent between languages (details discussed shortly). This strict requirement meant that a direct translation between languages was not often possible, but each sentence was semantically as similar as possible to its cross-linguistic equivalent. The sentences' syntactic structure naturally differed cross-linguistically; within languages, some syntactic variation was necessary to construct enough sentences which were prosodically identical and semantically equivalent between languages. Since syntax and (produced) rhythm may interact (see e.g. Classe 1939, Kohler 2009a, Lehiste 1980), this factor will be explored during the results analysis. All sentences were a nine-syllable Intonational Phrase (IP) comprising three smaller phrases, which (as in experiment 2) are termed 'rhythmic groups' (RGs). Each RG had a prominent syllable in the same position, which had to be RG-final in Fr and RG-medial in SG, according to each language's prosodic structure (details and references in chapter 2).

Fr has no lexical stress; the domain of prominence (i.e. RGs within IPs) usually corresponds to a syntagma such as a noun-, verb-, prepositional- or adverbial-phrase (Di Cristo 2000, Jun and Fougeron 2000). RG-final syllables are always prominent, marked by lengthening and often a rising f₀ IP-medially and falling f₀ IP-finally. Table 5-1 illustrates the prosodic

structure of the Fr sentences. Most were: [article + noun] + [auxiliary/modal + past participle/infinitive] + [NP/Prep-P/Adv-P], or the [Prep-P/Adv-P] came first.

RG	1			2			3		
Syllable	1	2	3	4	5	6	7	8	9
Prominence	X			X			X		
Example	L'en	sei	gnante	a	co	nnu	les	é	lèves
Gloss	The teacher			has known/knew			the pupils		

Table 5-1 – Prosodic structure of Fr stimuli

A native Parisian Fr speaker was recorded, and produced the expected pitch pattern on all sentences: a rise during syllables 3 and 6 and a fall during syllable 9. The original plan was to create one set of Fr and one set of SFr stimuli. Two native SFr speakers (Bern and Valais cantons) were recorded, who showed intra- and inter-speaker inconsistency in their production of f_0 rises. In most cases syllable 2 had a rise, but in RG 2, either syllable 5 or 6 had a rise; in RG 3, sometimes syllable 8 had a rise, and 9 was always falling. Miller (2007) also observed these variant pitch patterns in SFr, and noted the extreme inter- and intra-speaker variation across and within speech styles. Miller (2007) suggested that greater variation could result from greater familiarity of speaker and experimenter, and more spontaneous speech. Here neither speaker was well acquainted with me, and the sentences were read. (Their spontaneous speech, recorded from our conversation in (S)Fr beforehand, also revealed pitch-pattern variation.) The mixture of Fr and SFr pitch patterns in their recordings might have resulted from a psychological conflict. On the one hand, both speakers showed great willingness to participate, especially since they knew their ‘Swiss-ness’ was of interest, and they were explicitly encouraged to speak as they would in Switzerland. On the other hand, their shift towards a more Fr form might have resulted from a (subconscious) feeling of linguistic insecurity (as chapter 2 discussed, SFr may be deemed less socially prestigious than Fr, e.g. Singy 1996), perhaps related to the fact that my L2 Fr is based on the variety from France.

The inconsistent location of SFr rises made it impossible to construct stimuli with consistent regular prominence. When syllable 2/5/8 had a rise, this gave the impression of RG-penultimate (not RG-final) prominence, though it is unknown whether native speakers perceive these rises as more prominent than final lengthened syllables (cf. Miller 2007). Consequently, only (Parisian) Fr stimuli were created and presented to Fr and SFr subjects (no sentences contained words that are different or nonexistent in SFr). Therefore, the SFr subjects did not hear speech in their own accent, which was the original rationale behind recording SFr speakers. Nevertheless, these subjects (university students) have daily exposure to Fr through the media, and Fr is taught/spoken in education (Bayard and Jolivet 1984, Miller 2007, Singy 1996); SFr students in higher education often acquire a variety more similar to Fr than their contemporaries who left

education earlier (Knecht and Rubattel 1984). Subjects had to compare versions of (durationally-/tonally-varied) sentences spoken by one Fr speaker, so subjects were not asked to judge how SFr and Fr rhythmicality compare. It would be interesting if SFr and Fr subjects' responses differ, and this would add to perceptual evidence of between-variety prosodic variation. Interpretation of results may involve reference to sociolinguistic factors such as attitudes towards more/less prestigious varieties.

SG has lexical stress, which is assigned to the lexical root, often the word-initial syllable. A polysyllabic content word with initial lexical stress preceded by a non-prominent function word forms a RG with medial prominence, marked by rising f_0 and syllable lengthening; thus IPs comprise RGs that often correspond to a syntagma (Häsler et al. 2005). (This is consistent with Reese's (2007) model of SG prominence based on feet which are ideally disyllabic and trochaic but deviations occur.) Table 5-2 illustrates the prosodic structure of the SG sentences. RG-medial prominence (rather than RG-final in Fr) was necessary, since it would be extremely difficult to construct thirty different sentences comparable in syntax and lexical complexity to the Fr sentences, if RG-final syllables had to be monosyllabic prominent content words. Most sentences were: [article + noun] + [auxiliary/modal + NP/Prep-P/Adv-P] + [past participle/infinitive], or the [article + noun] and [Prep-P/Adv-P] were reversed. Only verbs with prefixes that never receive lexical stress appeared in RG 3.

RG	1			2			3		
Syllable	1	2	3	4	5	6	7	8	9
Prominence	X			X			X		
Example	De	Lee	rer	würt	dSchüe	ler	er	chä	ne
Gloss	The teacher			will the pupils			recognise		

Table 5-2 – Prosodic structure of SG stimuli

A native Zürich German speaker was recorded, and produced the expected pitch pattern (according to studies discussed in chapter 2) on all sentences: a rise on syllables 2 and 5, sometimes completely within the syllable, sometimes beginning late and peaking in the following syllable; a less noticeable rise on syllable 8 before a sharp fall on syllable 9 (the speaker still perceived 8 as prominent).

The SG speaker and Fr speaker (both male) were recorded, on separate occasions in Cambridge University Phonetics Laboratory's sound-attenuated booth, reading the thirty sentences in their language. They were unaware of the precise purpose of their recordings. For the first read-through they were instructed to speak at a pace comfortable and natural for them. To check that this 'normal' pace was contrastable to a slower and faster one, they had to increase and decrease the pace for the second and third read-through respectively. For two final read-throughs they returned to a normal pace. The equipment used was a *Marantz* PMD670 solid-state

recorder, and a low-noise condenser *Sennheiser* MKH40P48 microphone with a cardioid frequency response. The recording mode was set to 16 bit linear PCM, with a 44.1 kHz sample rate. The files were saved as .wav format, then transferred onto an *iMac* (Mac OS X.5) via a USB cable, and displayed in *Praat*.

5.4.2 Duration and f0 values chosen for manipulations

In the pilot, despite the stimuli having stylised f0 contours, listeners did not find the sentences particularly unnatural, according to their post-test questionnaire answers. It was ultimately decided, however, to use more natural stimuli, so the conclusions drawn from results would be generalisable to natural speech. In each pilot stimulus, only duration *or* f0 was manipulated, but to explore the interdependence of these cues, co-varied stimuli were necessary. For each sentence, nine stimuli were created (3 duration x 3 f0 conditions). Table 5-3 describes, for each condition, the manipulation made on the prominent syllable of the second RG (SG syllable 5, Fr syllable 6). As will be explained in §5.4.3, each F0_{Norm}/DUR_{Norm} stimulus underwent resynthesis like all the other stimuli.

		f0		
		F0 _{Low}	F0 _{Norm}	F0 _{High}
duration	DUR _{Short}	-35%ms, -3st	-35%ms, original f0	-35%ms, +3st
	DUR _{Norm}	original duration, -3st	original duration, original f0	original duration, +3st
	DUR _{Long}	+35%ms, -3st	+35%ms, original f0	+35%ms, +3st

Table 5-3 – Duration and/or f0 manipulation made in nine conditions

This experiment worked in semitones (st), because pitch perception does not correspond linearly to absolute decreases/increases in Hz (Lehiste 1970, Nolan 2003). The best known psychoacoustic scale which captures this non-linear relationship is the essentially logarithmic st scale (Nolan 2003); others (not available in *Praat*) include Bark, mels and ERB-rate. The f0 peak was decreased/increased, which was straightforward for Fr, as the peak always occurred during the prominent syllable. In SG, the peak sometimes occurred in the syllable following the perceptually prominent one. After much contemplation, it was decided to decrease/increase the f0 peak, regardless of location relative to the prominent syllable, because if the f0 at prominent syllables' right-edge had been manipulated, the late-ending rises would have changed shape, so the stimuli might have sounded odd and not equivalent to the Fr stimuli in which rises did not change shape.

Unlike pitch, there is apparently no evidence that length perception relates to physical duration on a non-linear scale. The durational just noticeable difference (JND) between a reference sound and another increases as the reference duration increases (Lehiste 1970: 11-12).

Here, the duration of the to-be-manipulated syllable had a large range (345.47ms for SG, 206.98ms for Fr) across the thirty sentences. A percentage decrease/increase was implemented to avoid the problem that an absolute increase/decrease would be much more perceptible for the shortest than the longest to-be-manipulated syllables. The whole syllable was shortened/lengthened, since languages may vary in the ratio that consonants and vowels lengthen when speech rate decreases or syllables are made prominent, so if just vowel duration had been manipulated, it might have been more natural for one language than the other. This seems particularly problematic since SG, but not Fr, has phonological vowel- and consonant-length contrasts. Moreover, the syllable is apparently the f_0 movement (i.e. pitch-accent) domain in both languages. In the recordings, f_0 rose across the whole of prominent syllables, including consonants (when voiceless, f_0 jumped up from the previous to the following vowel). Therefore, the duration manipulations concerned the same domain as the f_0 manipulations. SG late f_0 rises involved two syllables. However, only the prominent syllable was shortened/lengthened in all SG stimuli, because the late rises were displaced, rather than longer than the rises within one syllable, and if two syllables had been shortened/lengthened in some SG stimuli, this could have decreased naturalness and cross-linguistic equivalence.

The duration/ f_0 manipulation values, which were near the ends of the acceptable (i.e. rhythmical) range found in the pilot, were chosen so that the stimuli satisfied two criteria¹.

1. Each durationally-/tonally-deviant sentence (i.e. any DUR_{Short} , DUR_{Long} , $F0_{Low}$, $F0_{High}$ stimuli) should generally be perceptibly different from the non-deviant sentence (i.e. the $F0_{Norm}/DUR_{Norm}$ stimulus – also resynthesised). All differences between DUR_{Short} and DUR_{Norm} , and DUR_{Norm} and DUR_{Long} stimuli were above the reported JNDs of 10-40ms for speech sounds around 30-300ms (Lehiste 1970).
2. The durationally-/tonally-deviant sentences must not be so different from the $DUR_{Norm}/F0_{Norm}$ sentence that they sound impossibly unnatural, and thus listeners would at least contemplate choosing them in the rhythmicality judgment task.

A balance between these criteria was sought through a discrimination test.

5.4.2.1 AXB discrimination pre-test

For sixteen of the thirty sentences, nine stimuli per sentence were created in which, as described above, the medial-prominent syllable had manipulations of $\pm 35\%$ ms and/or ± 3 st. A

¹ For SG, it was also checked that phonologically-long vowels when shortened were considerably longer than their phonologically-short counterparts at DUR_{Norm} , and that phonologically-short vowels when lengthened were considerably shorter than their phonologically-long counterparts at DUR_{Norm} (means: long vowels at $DUR_{Short}=121.68$ ms; short vowels at $DUR_{Norm}=107.61$ ms; short vowels at $DUR_{Long}=145.27$ ms; long vowels at $DUR_{Norm}=187.20$ ms).

further nine stimuli per sentence were created, with manipulations of greater magnitude: $\pm 45\%$ ms and ± 5 st. In each trial, subjects were presented with three sentences, AXB, where A or B was identical to X, and the non-identical stimulus had the same words, but differed by one step of duration and/or f0 excursion manipulation. The three sentences were not played successively, as this would have caused poor performance or even made the task impossible, because remembering three sentences is cognitively demanding. Instead three buttons appeared on screen, labelled (left to right) ‘A’, ‘X’, ‘B’, which each played a stimulus when clicked (a maximum of three times). Subjects had to choose whether sentence X was identical to A or B.

The first block of sixteen trials presented the stimuli with $\pm 35\%$ ms/ ± 3 st manipulations. The second block tested the same sixteen manipulation combinations (e.g. F0_{Low}/DUR_{Short} versus F0_{Low}/DUR_{Norm}) except that the stimuli had the $\pm 45\%$ ms/ ± 5 st (i.e. more obvious) manipulations. Table 5-4 shows that the Fr subjects scored (non-significantly) lower than the SG subjects in the first block [$t(22)=-1.16$, $p>0.05$], but the scores in the second block were virtually identical. A possible factor in the cross-linguistically different results is that the manipulation (which distinguishes A/B from X) is RG-medial for SG, so is heard with a constant context either side within one RG, whereas for Fr the manipulation is RG-final, so the preceding syllable in is the same RG and the following syllable in the next RG.

Language (12 subjects per group)	First block (less obvious manipulations)	Second block (more obvious manipulations)
(S)Fr	70.31	89.06
SG	77.60	89.58

Table 5-4 – Mean percentage of correct responses in AXB task (chance=50%)

According to binomial probability, a subject needs to score $\geq 68.75\%$ in one block (sixteen trials) to discriminate above chance when $p=0.05$, and $\geq 87.50\%$ when $p=0.001$. Since responses to the less obvious manipulations were significantly above chance (when $p=0.05$), these stimuli were sufficiently distinguishable from each other to use in the main rhythmicity-judgement experiment, in which the smallest between-stimulus differences that listeners would hear were the differences tested in the AXB test. Since responses to the more obvious manipulations were highly significantly above chance ($p=0.001$), these stimuli might have sounded too obviously manipulated and unnatural.

5.4.3 Preparation of stimuli

The following procedure applied for each of the thirty recorded sentences. One version of the sentence was selected for manipulation which had a tonal and durational pattern that most clearly demonstrated the prosodic structure outlined in §5.4.1. Only the normal-paced recordings, those without hesitations/mistakes, were considered. This version was extracted from the

recording and saved as a separate file, which was opened in *Praat* (see Figure 5-2), and then converted into a ‘manipulation object’ (see Figure 5-3).

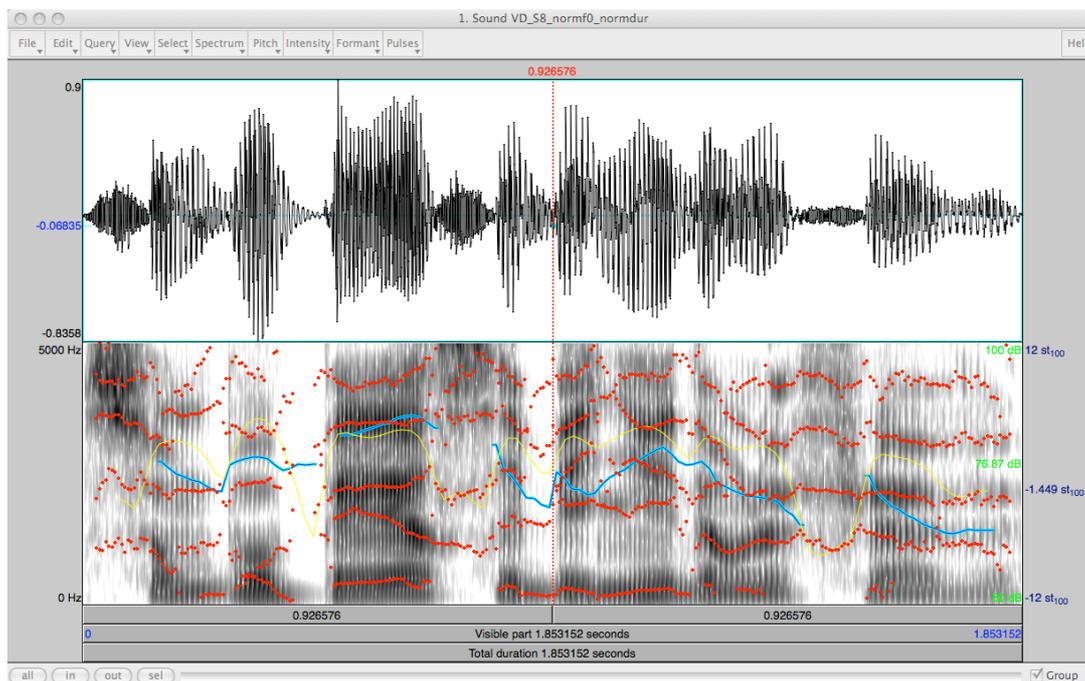


Figure 5-2 – Example sentence (Fr: *Son beau-père s'est baigné dans le fleuve*) displayed as spectrogram and waveform

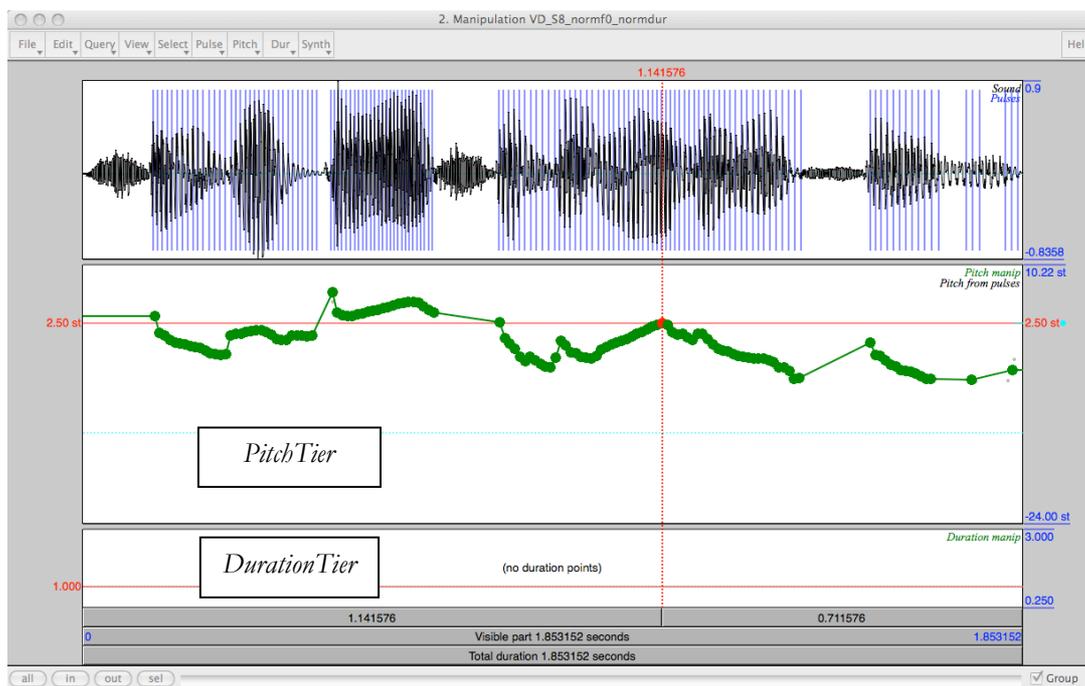


Figure 5-3 – Sentence in Figure 5-2 converted to ‘manipulation object’ with waveform, glottal pulses (blue), *PitchTier* and *DurationTier*

In the manipulation object, the original f_0 was displayed as a series of points joined by a line in the *PitchTier*. By default, these points appear at every glottal pulse (x-axis) with the appropriate st value (y-axis) (Figure 5-3). Then this f_0 contour was subjected to the ‘Stylise by

0.5st' command (for the algorithm, see Boersma 2005). The resulting f0 contour was not perceptibly stylised, but several points were removed, which aided manipulation of the f0 peak as there was always one clear peak point instead of a few tightly packed together (compare the red dot in Figures 5-3 and 5-4). The sentence was resynthesised and saved as the $F0_{Norm}/DUR_{Norm}$ stimulus. Subsequently the $F0_{High}/DUR_{Norm}$ and $F0_{Low}/DUR_{Norm}$ stimuli were made. The peak point was removed, then added again at exactly the same time position, but 3st higher or lower (Figure 5-4), and each sentence was resynthesised with the new f0 contour. These sentences and the *PitchTiers* were saved.

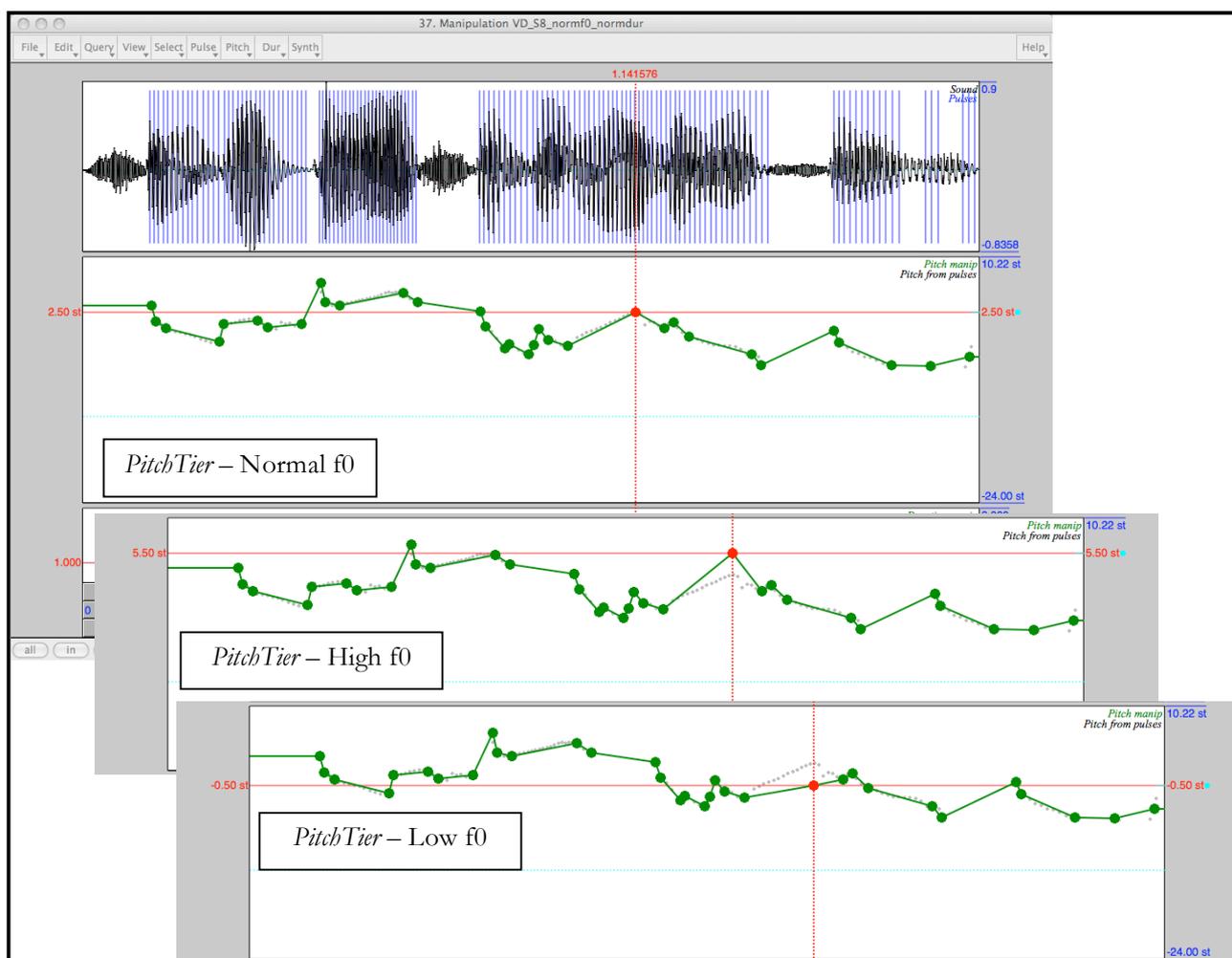


Figure 5-4 – *PitchTiers* showing the peak point (red dot) moved for the $F0_{High}$ and $F0_{Low}$ stimuli. Notice the original points at each glottal pulse in grey (below the green line for $F0_{High}$, above the green line for $F0_{Low}$)

Then the $F0_{Norm}/DUR_{Short}$ and $F0_{Norm}/DUR_{Long}$ stimuli were made. The previously saved *PitchTier* for $F0_{Norm}$ was added to the original manipulation object (i.e. the top *PitchTier* of Figure 5-4). A *Praat* script was written and run which created two *DurationTiers* (one each for shortening/lengthening), and which added points on these *DurationTiers* at the to-be-manipulated syllable boundaries, specifying that the sound between the points be shortened/lengthened by

35% (see Figure 5-5). Each *DurationTier* was added to the manipulation object one at a time, and each sentence was resynthesised with the new duration. These sentences and the *DurationTiers* were saved.

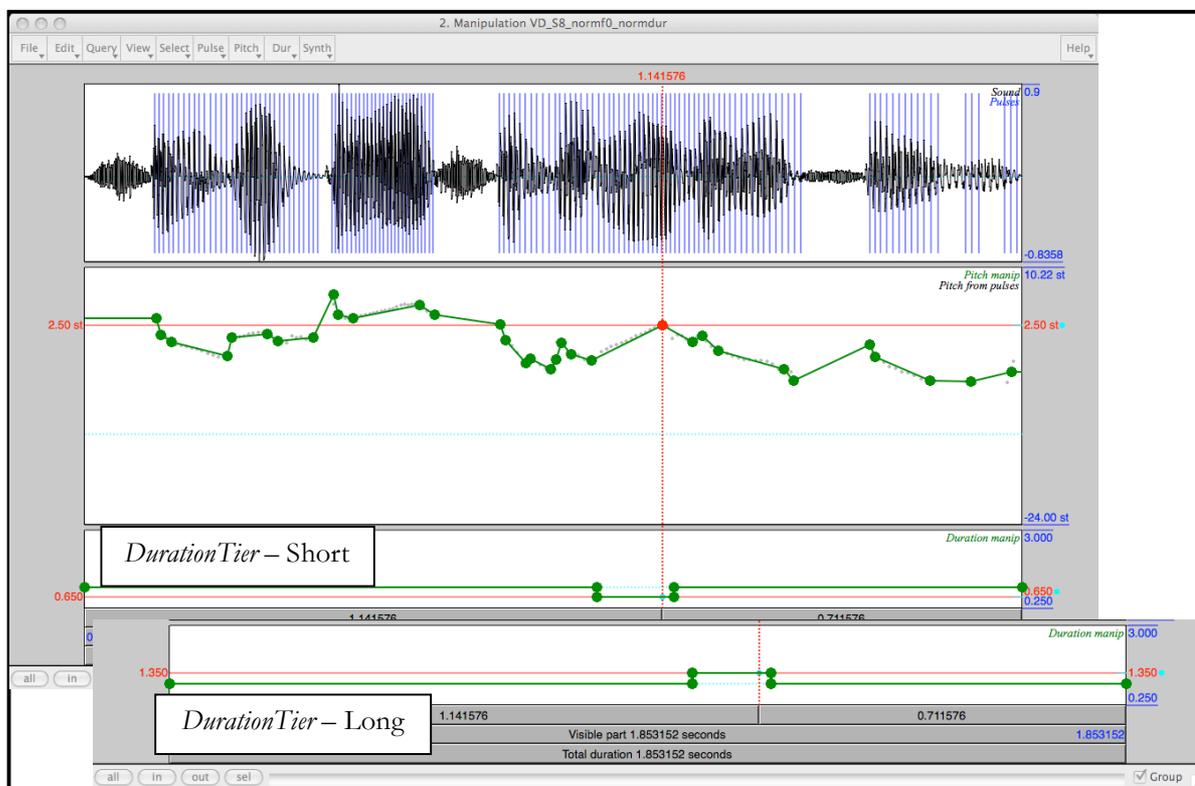


Figure 5-5 – *DurationTiers* with section specified to be shortened/lengthened shown as downward/upward indentation to green line

Then the remaining stimuli were created: $F0_{Low}/DUR_{Short}$, $F0_{Low}/DUR_{Long}$, $F0_{High}/DUR_{Short}$, $F0_{High}/DUR_{Long}$. The appropriate *PitchTier* and *DurationTier*, previously saved, were added to the original manipulation object, then each sentence was resynthesised with the appropriate f_0 and duration manipulation, and was saved.

5.5 Hypothesis

Now that the stimuli have been described, they help to illustrate the prediction. There is evidence that f_0 and duration are interdependent in the perception of isolated syllables (chapter 3) and rhythmic groups (chapter 4), and in the perceived rhythmicity of linguistically meaningless utterances (Barry et al. 2009, Grover and Terken 1995). Here, this interdependence is investigated with meaningful sentences. For each sentence, subjects must compare nine different stimuli (Table 5-5) and decide which one is most rhythmic.

	f0		
duration	F0 _{Low} /DUR _{Short}	F0 _{Norm} /DUR _{Short}	F0 _{High} /DUR _{Short}
	F0 _{Low} /DUR _{Norm}	F0 _{Norm} /DUR _{Norm}	F0 _{High} /DUR _{Norm}
	F0 _{Low} /DUR _{Long}	F0 _{Norm} /DUR _{Long}	F0 _{High} /DUR _{Long}

Table 5-5 – Nine stimulus conditions

If the F0_{Norm}/DUR_{Norm} stimulus were consistently preferred for every sentence, we could illustrate this as in Figure 5-6.

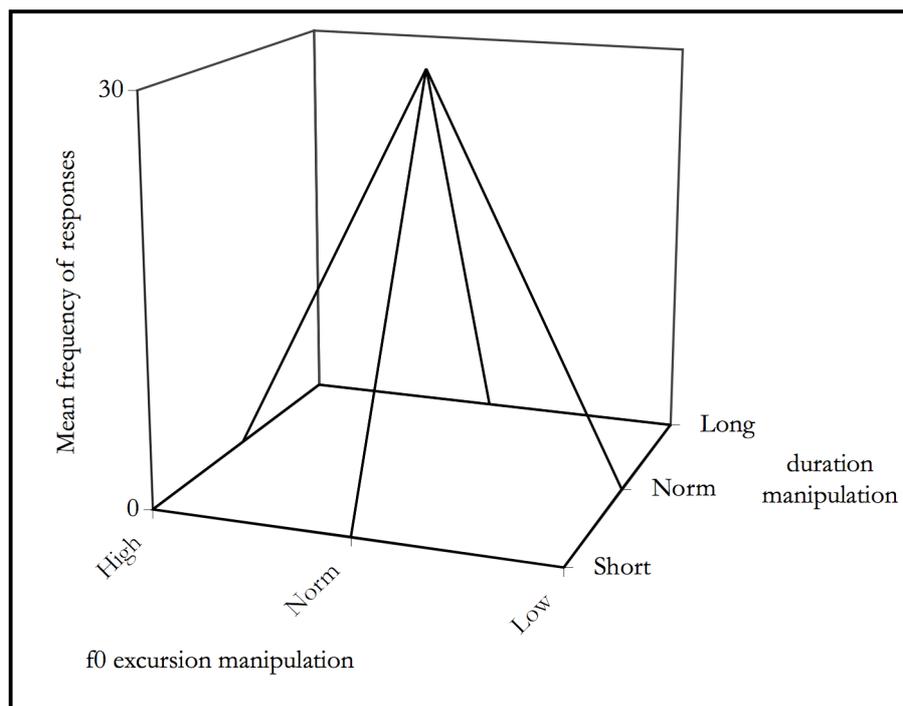


Figure 5-6 – Hypothetical data: if F0_{Norm}/DUR_{Norm} stimulus were preferred for all 30 sentences

However, if subjects prefer prominent syllables to be tonally non-deviant, but are more tolerant of durational deviance, we could illustrate this as in Figure 5-7; or, if they prefer prominent syllables to be durationally non-deviant, but are more tolerant of tonal deviance, we could illustrate this as in Figure 5-8. In both cases, subjects would probably still choose F0_{Norm}/DUR_{Norm} most often, but the frequency of responses for other stimuli would increase.

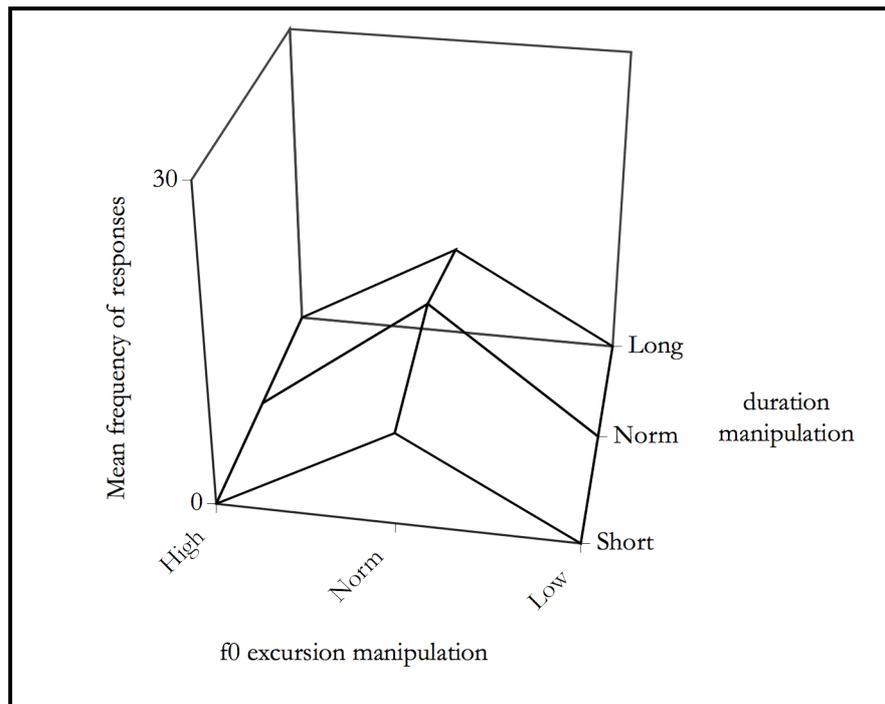


Figure 5-7 – Hypothetical data: if subjects prefer non-deviant f_0 , but are more tolerant of durational deviance

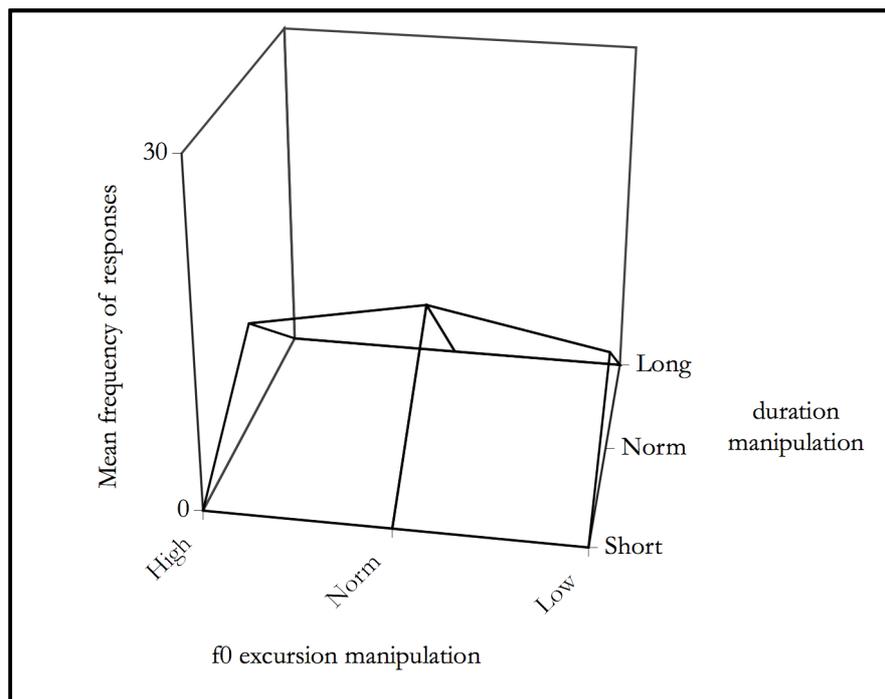


Figure 5-8 – Hypothetical data: if subjects prefer non-deviant duration, but are more tolerant of tonal deviance

There may be various reasons why subjects would sometimes prefer stimuli other than $F0_{\text{Norm}}/DUR_{\text{Norm}}$. If we consider experiment 1 (chapter 3), $F0_{\text{High}}/DUR_{\text{Short}}$ might be chosen relatively often, because the large f_0 excursion may increase the perceived length, thus

compensating perceptually for short physical duration. In experiment 2 (chapter 4), dynamic f0 was a more effective RG-boundary cue than increased duration for SG, but the reverse for (S)Fr, which suggests that we might observe cross-linguistic variation in subjects' tolerance of tonal/durational deviance here. However, these suggestions are speculative, and the present experiment is exploratory, as no evidence exists for exactly how f0 and duration are interdependent in the perceived rhythmicity of meaningful sentences.

It is predicted that the duration and f0 excursion of prominent syllables must both be appropriate for a sentence to sound rhythmic, but subjects' tolerance level for deviance of each cue might vary between languages, individuals and sentences. If so, subjects will choose the non-deviant $F0_{Norm}/DUR_{Norm}$ stimulus as most rhythmic most often, but they may sometimes choose other stimuli, i.e. the results will resemble Figures 5-7/5-8 more than Figure 5-6.

5.6 Main experiment (A): preliminary test

In the pilot (§5.3), the 'adjustment' task had proved to be a feasible means of obtaining rhythmicity judgments. Listeners clicked around a circle of boxes that each played a sentence, until they heard the sentence they deemed the most rhythmic. Duration or f0 excursion of one syllable increased linearly when listeners clicked clockwise, by analogy with a volume-control knob increasing amplitude when adjusted clockwise. For this main experiment, the stimulus design was modified so that duration and f0 manipulations were co-varied. Therefore, the analogy with a knob relating to one linear acoustic change was no longer possible. An alternative design could have been to group stimuli into threes around the circle, such that when clicking clockwise, pitch would change fairly slowly whereas length would change up and down repeatedly. This was trialled with a few subjects, but their feedback showed that this was confusing. After contemplation, it was decided to drop the circle in favour of a 3x3 matrix of boxes in which the nine stimuli were randomised, to exclude the possibility that an ordering artefact might affect results. To confirm that this randomised design with stimuli less stylised than in the pilot was still feasible, a preliminary test of the main experiment was run, with twelve SG subjects (various dialects) and twelve (S)Fr subjects (seven Fr, five SFr: at this stage between-variety differences were not explored). In the same testing session, these twenty-four subjects participated in the AXB discrimination task (§5.4.2.1), with a break before the rhythmicity judgement task.

5.6.1 Procedure

Subjects sat at a *MacBook* laptop (Mac OS X.5) and listened through binaural *Sennheiser* HD555 headphones. The experiment was scripted and run in *Praat*. There were two identical versions, one with SG stimuli and German on-screen text, one with Fr stimuli and on-screen text. Before testing began, subjects read instructions in their native language (appendix 8.3.2;

German for SGs, Fr for (S)Fr(s)), and were given chance to ask for clarification. For each trial, the following question appeared on screen: ‘Which sentence has the most natural rhythm?’ The question in the pilot was: ‘Which sentence is the most rhythmical?’ Some subjects’ feedback suggested that rephrasing to ‘natural rhythm’ would make it clearer and less technical. The two questions have potentially different implications. In principle, a certain duration/f₀ manipulation might make a sentence sound more rhythmical, in the sense of more closely resembling a line of metrical verse, but this rhythm would probably not sound natural in conversational speech (cf. Fowler 1979: 378, Grover and Terken 1995: 30). This experiment, to make its results generalisable to real speech perception, concerns utterances that could communicate information in everyday speech. Therefore, ‘natural rhythm’ was more appropriate than ‘rhythmical’, which could have implied that subjects should base their judgements on a technical speech style with limited use.

Below the question appeared a 3x3 matrix of boxes. When clicked (a maximum of three times), each box played one of the nine stimuli (Table 5-5) preceded by 500ms of silence. In each trial, the experiment script randomly assigned one stimulus to each box. Subjects were free to click around the matrix for as long as and in the order they wanted. After they identified their preferred stimulus, they clicked a ‘submit response’ button below the matrix. There were eighteen trials (i.e. a sub-set of the thirty sentences, as this preliminary test needed to confirm that this task was feasible), including one practice trial, and a short break after every six. Subjects could ask for clarification after the practice, though none did. The whole experiment lasted approximately thirty minutes. Subjects then filled in a post-test questionnaire to rate task difficulty.

5.6.2 Preliminary test: results

Figure 5-9 shows how often subjects judged each stimulus type as most rhythmically-natural throughout the eighteen trials. The same frequency scale (0–7) is used for both graphs, for easy comparison of the pattern.

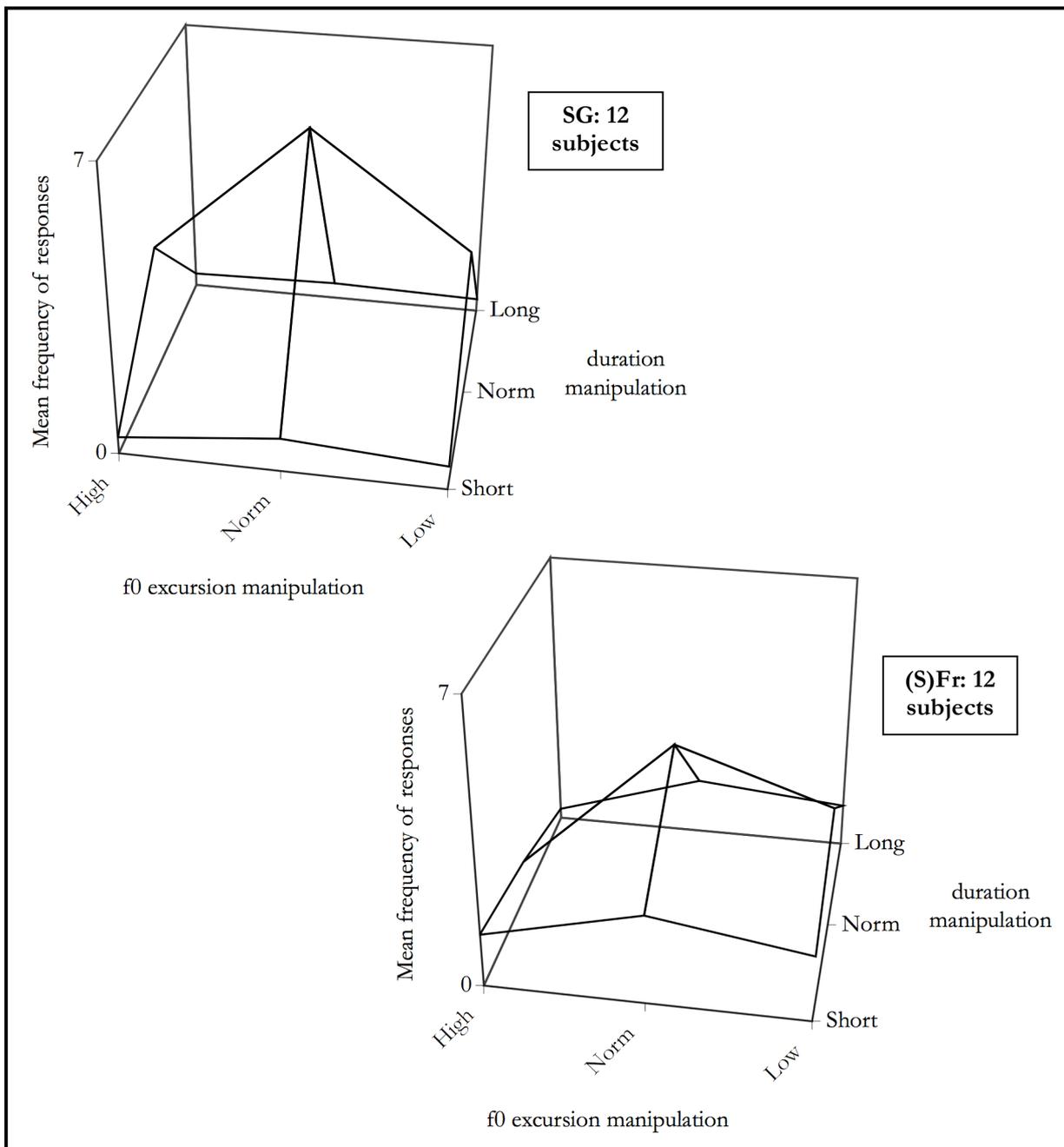


Figure 5-9 – Preliminary test: mean (across subjects) frequency of responses per stimulus type (out of 18 sentences)

The SG graph resembles Figure 5-8 from the predictions (§5.5): there is a peak at $F0_{Norm}/DUR_{Norm}$, and subjects often tolerated a deviant f0 excursion but hardly ever tolerated a deviant duration. The (S)Fr graph is more complex: there is a peak at $F0_{Norm}/DUR_{Norm}$, and subjects often tolerated a shorter deviant duration and a lower deviant f0 excursion. These results look interesting, but are not discussed in detail, because only eighteen sentences were included and only twelve subjects per language participated, who also did the AXB test, which may have drawn their attention to the acoustic manipulations. Moreover, the SGs were from various cantons (see later discussion). (Yet these factors turned out to have no influence, since Figure 5-9

resembles the final results in Figure 5-10.) The post-test questionnaire confirmed that subjects found the matrix design no harder than the previous circle presentation, so this preliminary test served its purpose.

5.7 Main experiment (B): final version

5.7.1 Subjects

All subjects (Table 5-6) reported normal hearing and were monolingual, i.e. had not learned another language before obligatory foreign-language classes starting around 9-11 years.

Language	Total	Age (years)		Sex	
		Range	Mean	Male	Female
SG	47	20–37	25.8	14	33
		18–37	21.9	7	43
SFr	50	18–38	25.8	18	30

Table 5-6 – Summary of subjects

The locations where each subject group was recruited and tested were as in experiment 1 (chapter 3). The SGs, mostly university students, were all life-long residents of the Zürich canton. If subjects had been included who had lived elsewhere, and so had a somewhat mixed dialect (see chapter 2), they might have been distracted by syntactic/lexical differences between the Zürich German stimuli and other dialects familiar to them. This might have led them to judge rhythmicity based on factors other than prosodic manipulations, or perhaps be less accurate generally in judging rhythmicity. The SFr, mostly university students, were from various SFr-speaking cantons (many from Neuchâtel). Regional variation in SFr accents is minor compared to SG dialectal variation (see chapter 2), and in any case, SFr listened to Fr stimuli.

Many Frs were university students in Cambridge, and some were briefly visiting Cambridge; none had lived in the UK before age 18. Thirty-three grew up in northern France, and fifteen in southern France²: the areas above and below a line roughly from Bordeaux (west) to Lyon (east). This broad north-south imaginary divide is often cited in illustrating Fr regional variation (mainly segmental) in some speakers' pronunciation (e.g. Ball 1997, Carton et al. 1983, Hawkins 1993), though Parisian Fr is the socially prestigious accent (Gadet 2007, Lodge 1993, Walter 1988). According to anecdote, southern accents are 'chantant' ('singing') (Ball 1997: 87). In an accent evaluation experiment, Carton (1987) found that subjects with no Fr knowledge

²The few who had moved around within France were classified according to where they spent the majority of childhood.

judged southern accents to be more ‘lively’ than northern accents. Phonetic analyses of the stimuli showed no north-south difference in intonation, but in the southern accents non-prominent syllables were significantly more irregularly timed. Therefore, Carton (1987: 37-38) suggested, the rhythm may have given this impression of ‘liveliness’. This could have resulted from the possibility that southern speakers may pronounce more optional schwas than northern speakers (Hawkins 1993), and may insert schwas where they would never appear in northern Fr (Carton et al. 1983). Experiments 1 and 2 (chapters 3-4) did not split Fr subjects by region, because those stimuli had stylised f0 and included no schwas³. This experiment will explore whether southern subjects, who may have grown up hearing/speaking a prosody subtly different from Parisian Fr, might respond differently from northern subjects now that the stimuli are more prosodically complex.

5.7.2 Procedure

This experiment was the same as in part A (§5.6), except that all thirty sentences were used, so there were thirty main trials with a short break after every seven or eight, and one practice trial. The experiment lasted approximately forty-five minutes. Subjects were assigned to one of three groups, each of which was presented with a different randomised order of the thirty sentences. The task instructions warned subjects about a post-test questionnaire (appendix 8.3.3). This asked subjects to rate task difficulty, describe their judgment criteria and what rhythm meant to them, detail their technique for the task, comment on anything unusual in the speaker’s voice, and list their musical training. This information was collected since it might give insight into intra-/inter-subject variability in rhythmicality judgements.

5.7.3 Results

The following sections report the analysis of responses to stimuli (including intra-/inter-subject consistency), followed by the analysis of the post-test questionnaire responses.

5.7.3.1 Rhythmicality judgements

Figure 5-10 shows how often subjects judged each stimulus type as most rhythmically-natural throughout the thirty trials. The same frequency scale (0–9) is used for all graphs, for easy comparison of the pattern. We see that SGs often tolerated a deviant f0 excursion of the medial-prominent syllable but hardly ever tolerated a deviant duration, whereas (S)FrS often tolerated a shorter duration and a lower f0 excursion.

³ When experiment 2’s Fr data (chapter 4) were subjected to a two-way mixed-measures ANOVA with the factors *region* (N, S) and *cue(s)* (x7), no main effect of *region* occurred [$F(1,34)=2.321, p>0.05$], nor an interaction of *region* × *cue(s)* [$F(3.839,130.535)=0.469, p>0.05$].

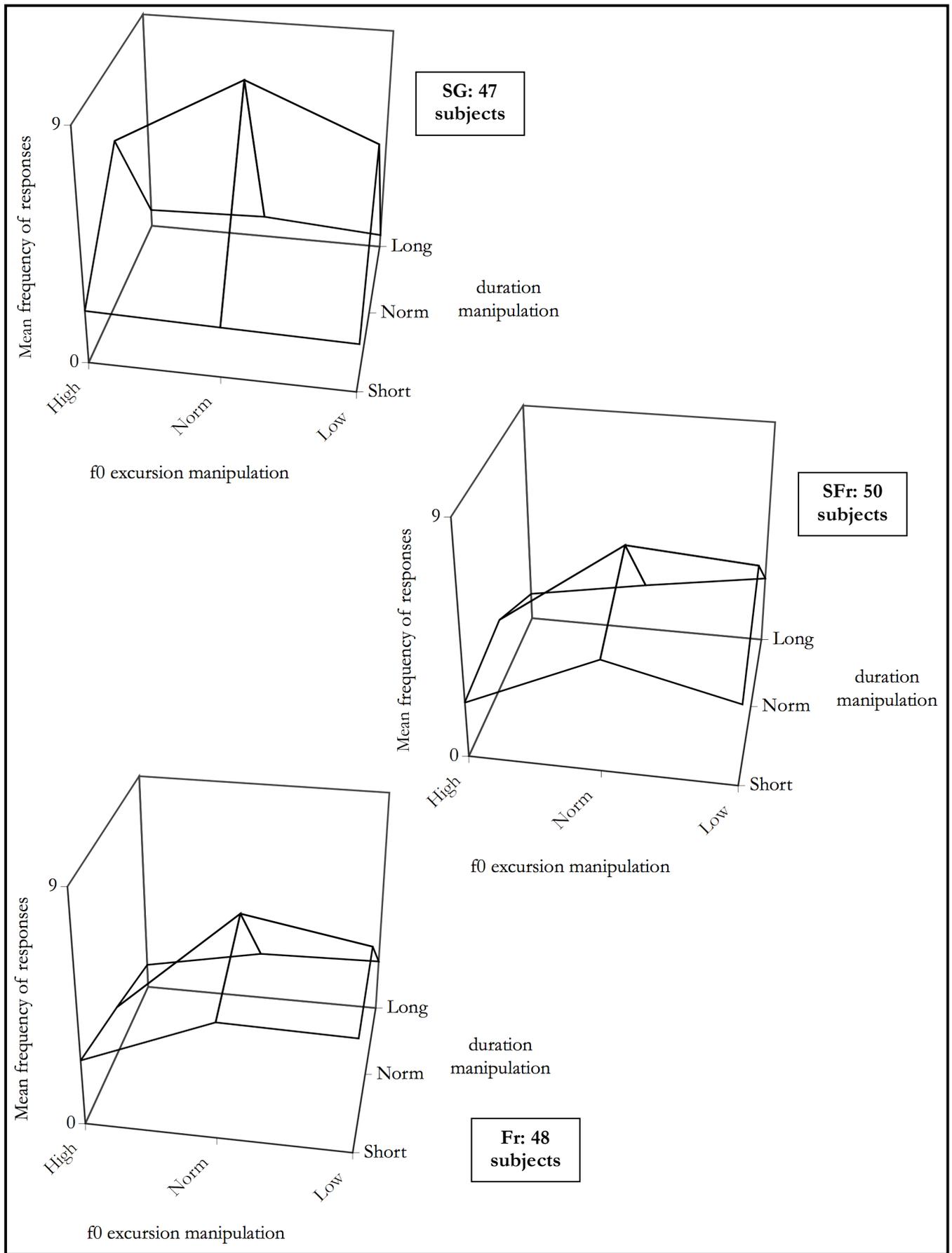


Figure 5-10 – Mean (across subjects) frequency of responses per stimulus type (out of 30 sentences)

The graphs reveal differences for stimulus types and languages, which were analysed statistically with an ANOVA (in *SPSS*). Beforehand, the response-frequency data were explored for normality and homogeneity of variance. Most data-series were relatively normally distributed according to visual inspection of histograms, though some were significantly non-normal according to Shapiro-Wilk tests and skewness/kurtosis statistics ($p < 0.05$) (see Field 2005). ANOVA is robust against violation of the normality assumption, but it is important that variances are homogeneous if sample sizes are unequal (Howell 2007), as they were here (SG=47; SFr=50; Fr=48). According to Levene tests, one data-series ($F_{0\text{Low}}/DUR_{\text{Long}}$) significantly violated the homogeneity of variance assumption ($p < 0.05$). A square-root transformation was applied to all data, and subsequent Levene tests revealed no violations ($p > 0.05$). The transformed data were input to a three-way ANOVA with the factors *f0 excursion* ($F_{0\text{Low}}$, $F_{0\text{Norm}}$, $F_{0\text{High}}$), *duration* (DUR_{Short} , DUR_{Norm} , DUR_{Long}) (both repeated-measures) and *language* (SG, SFr, Fr). All main effects and interactions were highly significant (Table 5-7).

Source	df	Mean square	F	p
f0 excursion	2	38.063	55.668	<0.0001**
f0 excursion × language	4	8.708	12.736	<0.0001**
Error	284			
duration	1.806	183.784	15.520	<0.0001**
duration × language	3.613	22.511	18.927	<0.0001**
Error	256.494			
f0 excursion × duration	4	2.705	7.870	<0.0001**
f0 excursion × duration × language	8	1.189	3.460	<0.01*
Error	568			
language	2	0.524	17.716	<0.0001**
Error	142	0.030		
<ul style="list-style-type: none"> • A Greenhouse-Geisser correction was applied since sphericity could not be assumed (Mauchly's test, $p < 0.001$) • significance: ** $p < 0.0001$; * $p < 0.01$ 				

Table 5-7 – ANOVA output: *f0 excursion* × *duration* × *language*

Post-hoc tests for *language* ('Gabriel', since sample sizes were unequal; Field 2005) revealed a significant difference between SG and SFr, SG and Fr, but not SFr and Fr. When the interactions were explored, a pattern emerged which is best interpreted with reference to Figure 5-10 (above) and Figure 5-11 (mean responses across three languages).

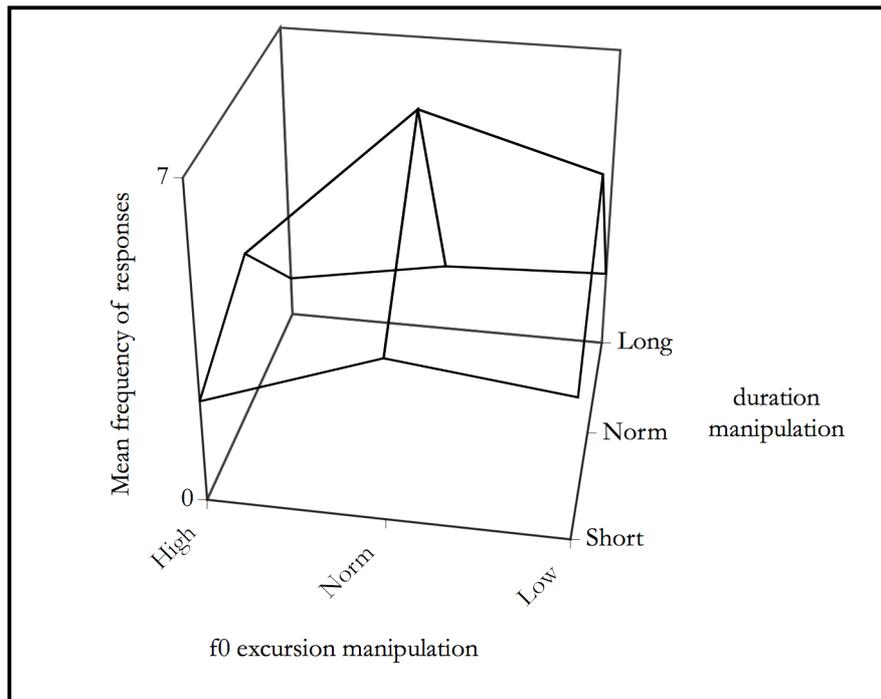


Figure 5-11 – Mean (across all subjects: 47+50+48=145) frequency of responses per stimulus type (out of 30 sentences)

Imagine the graph as an (imperfect) ridge tent. The line for responses at DUR_{Norm} represents the (slanted) central horizontal ridge pole, running left to right across three upright supporting poles (not visible) representing responses at $F0_{High}$, $F0_{Norm}$ and $F0_{Low}$ respectively. The lines for responses at DUR_{Short} and DUR_{Long} represent the lower edges of the canvas, to which three guy ropes (not visible) are attached on each side at points aligned with the upright poles, again representing $F0_{High}$, $F0_{Norm}$ and $F0_{Low}$. The right-hand entrance is at $F0_{Low}$ and the left-hand entrance at $F0_{High}$. The lines joining the nine stimulus-type data-points do not represent continuous data between the categorical stimulus types.

First, consider the $f0 \times duration$ interaction with all languages together (Figure 5-11). The horizontal ridge pole is more steeply slanted than the lower edges of the canvas: if the medial-prominent syllable had non-deviant duration (DUR_{Norm}), the $f0$ excursion of that syllable ($F0_{Low}$, $F0_{Norm}$ or $F0_{High}$) was more important in subjects' rhythmicity judgements than if the medial-prominent syllable had deviant duration (DUR_{Short} or DUR_{Long}). Likewise the right- and left-hand entrances are lower than the height of the tent at the centre: if the medial-prominent syllable had non-deviant $f0$ ($F0_{Norm}$), the duration of that syllable (DUR_{Short} , DUR_{Norm} or DUR_{Long}) was more important in subjects' rhythmicity judgements than if the medial-prominent syllable had deviant $f0$ excursion ($F0_{Low}$ or $F0_{High}$).

This general pattern is subtly different cross-linguistically (Figure 5-10), i.e. the $f0 \times duration \times language$ interaction in which SG and (S)Fr differed significantly. (S)Fr and Fr were not significantly different, though it is noticeable that Frs more often than SFrs perceived the most

natural-sounding rhythm if the medial-prominent syllable was $F0_{Low}/DUR_{Short}$.) The SG tent is tall and steep-sided, whereas the (S)Fr tents are lower and flatter: the difference between durationally-deviant (DUR_{Short} , DUR_{Long}) and durationally-non-deviant (DUR_{Norm}) stimuli, in terms of the importance of the $f0$ of these stimuli in subjects' rhythmicity judgements, was greater for SG than (S)Fr. The SG tent's sides are about equally steep and straight-edged, whereas the (S)Fr tents' front-facing side is flatter and less straight-edged than the back-facing side: in SG, there was little difference between DUR_{Short} and DUR_{Long} stimuli, in terms of the importance of the $f0$ of these stimuli in subjects' rhythmicity judgements, but in (S)Fr, the $f0$ was more important in judgements for DUR_{Short} than DUR_{Long} stimuli. The SG tent's right-hand and left-hand entrances are similarly high, whereas the (S)Fr tents' right-hand entrance is higher than the left-hand (which is almost nonexistent): in SG, there was little difference between $F0_{Low}$ and $F0_{High}$ stimuli, in terms of the importance of the duration of these stimuli in subjects' rhythmicity judgements, but in (S)Fr, the duration was more important in judgements for $F0_{Low}$ than $F0_{High}$ stimuli.

To explore whether subtle prosodic variation between northern and southern Fr accents (reported by e.g. Carton et al. 1983) influenced responses, the Fr data were divided by region (north/south: Figure 5-12). We see that the two groups differ slightly, particularly at $F0_{Low}/DUR_{Short}$, where the southern Fr responses are more similar to SFr (compare with Figure 5-10) than the northern Fr responses are.

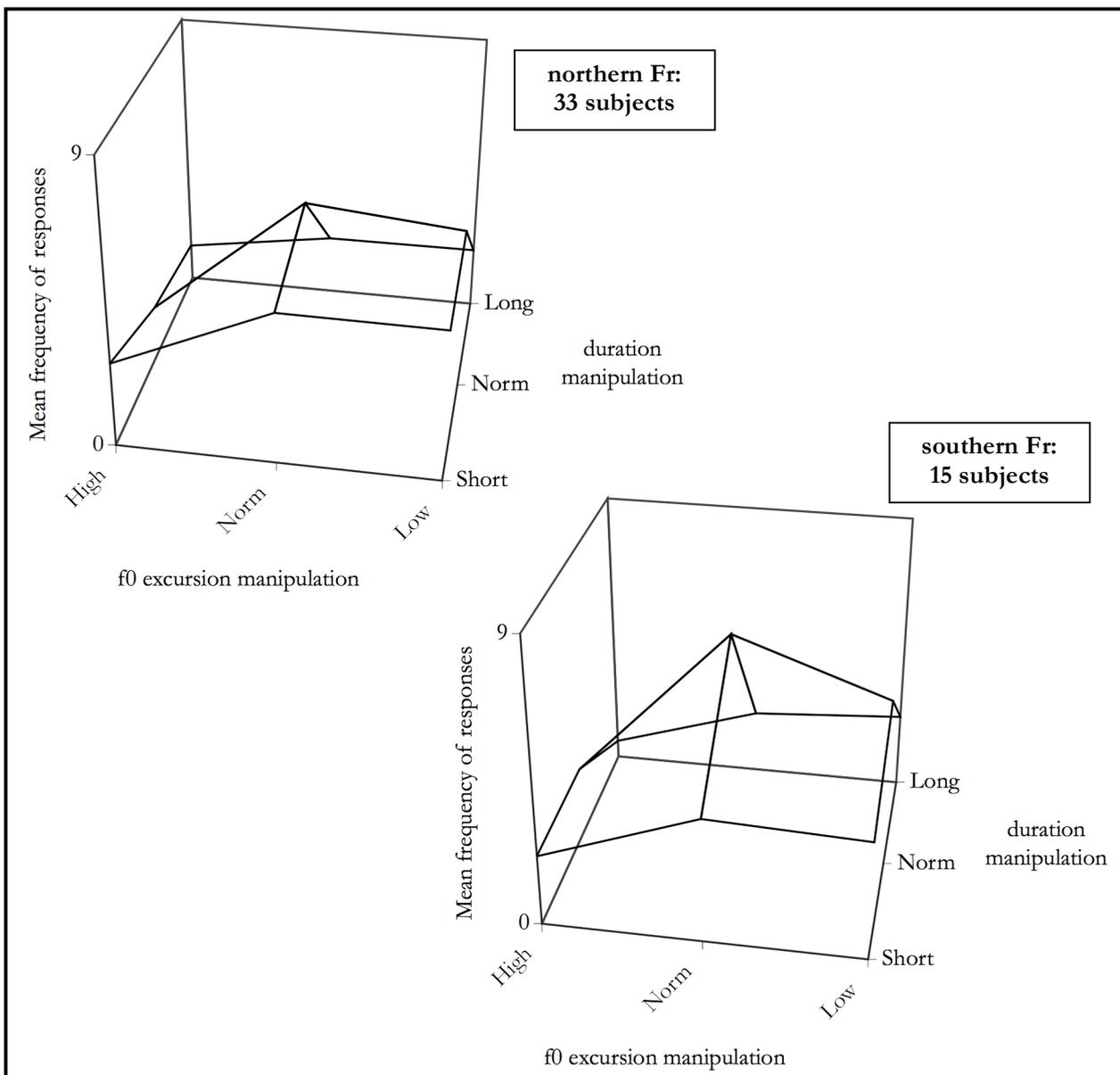


Figure 5-12 – Mean (across subjects) frequency of responses per stimulus type (out of 30 sentences)

An ANOVA was run on the Fr response-frequency data (untransformed), with the factors *f0 excursion* (F_{0Low} , F_{0Norm} , F_{0High}), *duration* (DUR_{Short} , DUR_{Norm} , DUR_{Long}) (both repeated-measures) and *region* (N, S). Since sample sizes were unequal (33 N, 15 S), a model with type III sums of squares (*SPSS* default) and data with homogenous variances were imperative, which Levene tests demonstrated was the case. No main effect of, or interactions with, *region* occurred (Table 5-8), and all other significant effects accord with those in Table 5-7. Fr rhythmicity judgements were not affected by subjects' regional origin, just like SFr and Fr judgements did not differ significantly overall.

Source	df	Mean square	F	p
f0 excursion	2	224.493	28.082	<0.0001**
f0 excursion × region	2	1.576	0.197	0.821
Error	92	7.994		
duration	1.702	315.749	26.387	<0.0001**
duration × region	1.702	22.397	1.872	0.166
Error	78.313	11.966		
f0 excursion × duration	4	31.447	7.702	<0.0001**
f0 excursion × duration × region	4	3.155	0.773	0.544
Error	184	4.083		
region	1	< 0.0001	< 0.0001	1.000
Error	46	< 0.0001		
<ul style="list-style-type: none"> • A Greenhouse-Geisser correction was applied since sphericity could not be assumed (Mauchly's test, $p < 0.05$) • significance: ** $p < 0.0001$ 				

Table 5-8 – ANOVA output: *f0 excursion × duration × region*

5.7.3.2 Intra- and inter-subject consistency

The response data were then examined for individual subjects and compared between individuals and languages. For each individual, the standard deviation (sd) of response frequency across stimulus types was calculated. The lower the sd, the more evenly-spread their thirty responses were over the nine stimulus types, i.e. a flatter tent, to use the above analogy. We see from Figure 5-13 that 70% of SGs had a sd above 3, and over 10% above 5, whereas over 50% of (S)Fr had a sd below 3, and nearly 10% below 2. SG individuals perceived one or two prosodic manipulations as most rhythmically-natural more consistently than (S)Fr individuals, who showed a wider spread of preferred stimulus types.

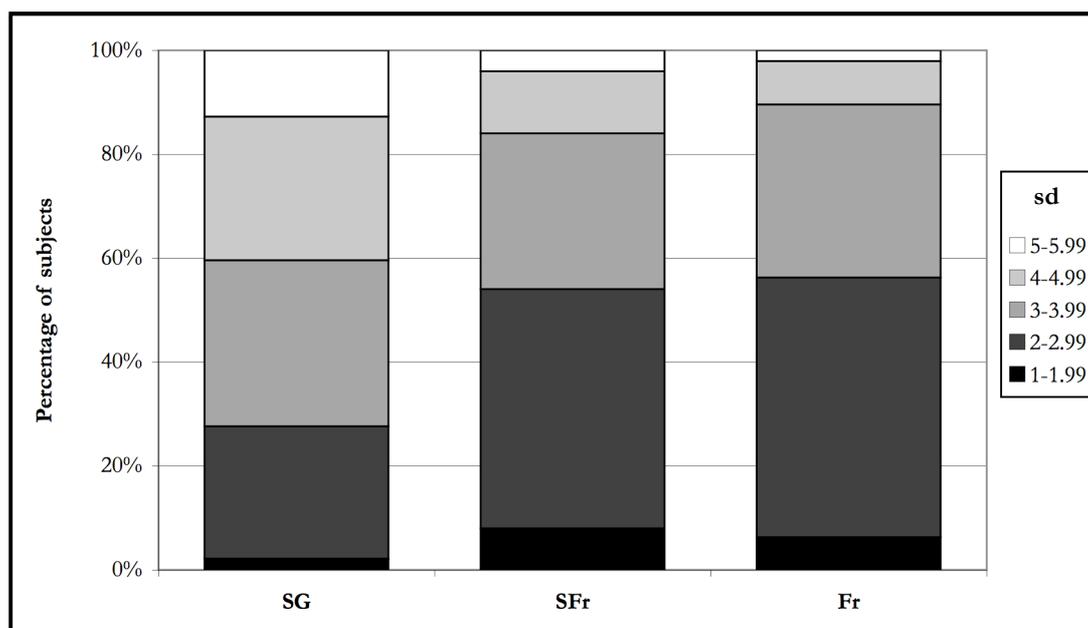


Figure 5-13 – Percentage of subjects whose sd of response frequency across stimulus types fell into each shaded range

This intra-subject variation might result from the fact that the thirty sentences, though prosodically identical, had different words, and that nearly two thirds of the sentences per language had one of two syntactic structures, whilst the other third differed slightly in syntax (for discussion of rhythm and syntax, see e.g. Classe 1939, Lehiste 1980). The response-frequency data per stimulus type is shown for individual sentences in Figure 5-14. These graphs are not designed to allow comparison of responses between one individual sentence and another. Rather the graphs illustrate that overall there was relatively high between-sentence variation in terms of how many subjects perceived each stimulus type to be most rhythmically-natural for that sentence, more so for (S)Fr than SG.

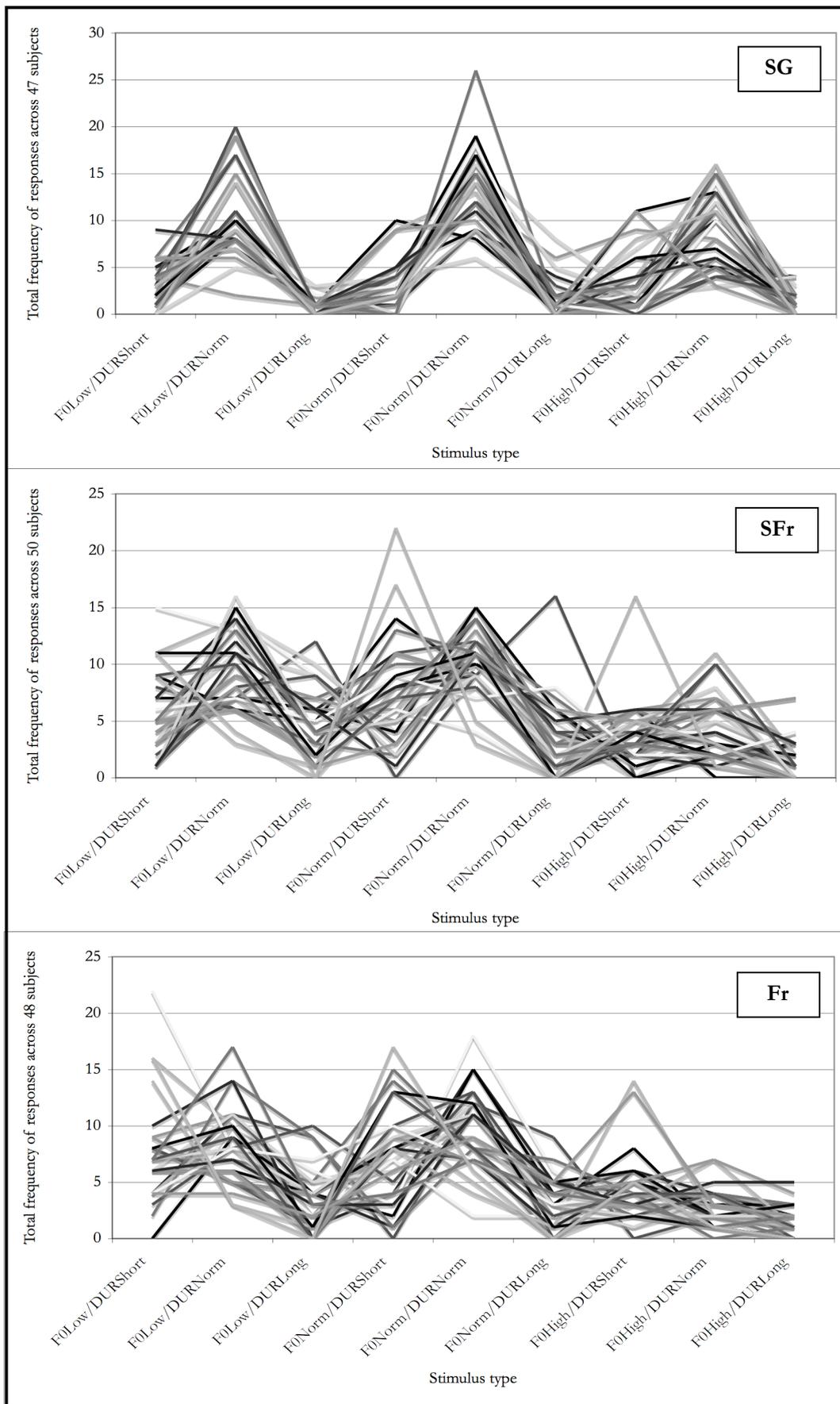


Figure 5-14 – Total frequency of responses per stimulus type; separate line per sentence

To test for a possible effect of sentence and syntax on responses, these data (total response frequency per stimulus type per sentence) were input to a linear mixed model using the R software environment. This model was like a two-way ANOVA for *stimulus type* and *language* (fixed factors), but included the random factors *sentence* and *syntactic structure*, which are random variables since these sentences were sampled randomly from all the possible sentences and structures in each language and were not the factor manipulated for investigation (Baayen 2008). To test for an interaction of *sentence* and *language*, the model allowed the slope of the effect of *language* to vary across *sentences* (see Baayen 2008). Table 5-9 shows that the variance of each random factor was almost 0, so these variables are superfluous to the model (Baayen 2008, Faraway 2006); since 0 falls within the lower and upper bounds of the confidence interval for the correlations between *sentence* and *language*, this interaction is also superfluous to the model (Baayen 2008)⁴. That is, the lexical and syntactic differences between sentences did not affect rhythmicity judgements overall, nor account for the cross-linguistic variation.

Random factor	Variance	Correlations		
		mean	lower bound (95% confidence)	upper bound (95% confidence)
sentence	Intercept	7.4496e-09	–	–
	SFr	5.0598e-09	–0.06681	–0.5684
	SG	4.9967e-09	–0.05843	–0.5495
syntactic structure	Intercept	4.9964e-09	–	–

Table 5-9 – Output of regression model (random factors only)

Three-dimensional Figure 5-10 would have been visually complicated if it had included lines illustrating statistics of inter-subject variation. These are demonstrated in Figure 5-15 by plotting two-dimensionally the data in Figure 5-10.

⁴ For another experiment (reported in chapter 6), the data from this experiment were input to logistic regression analyses (generalised linear mixed models) which initially included the random factors *sentence* and *syntactic structure*, but these were removed since they were likewise found to be superfluous.

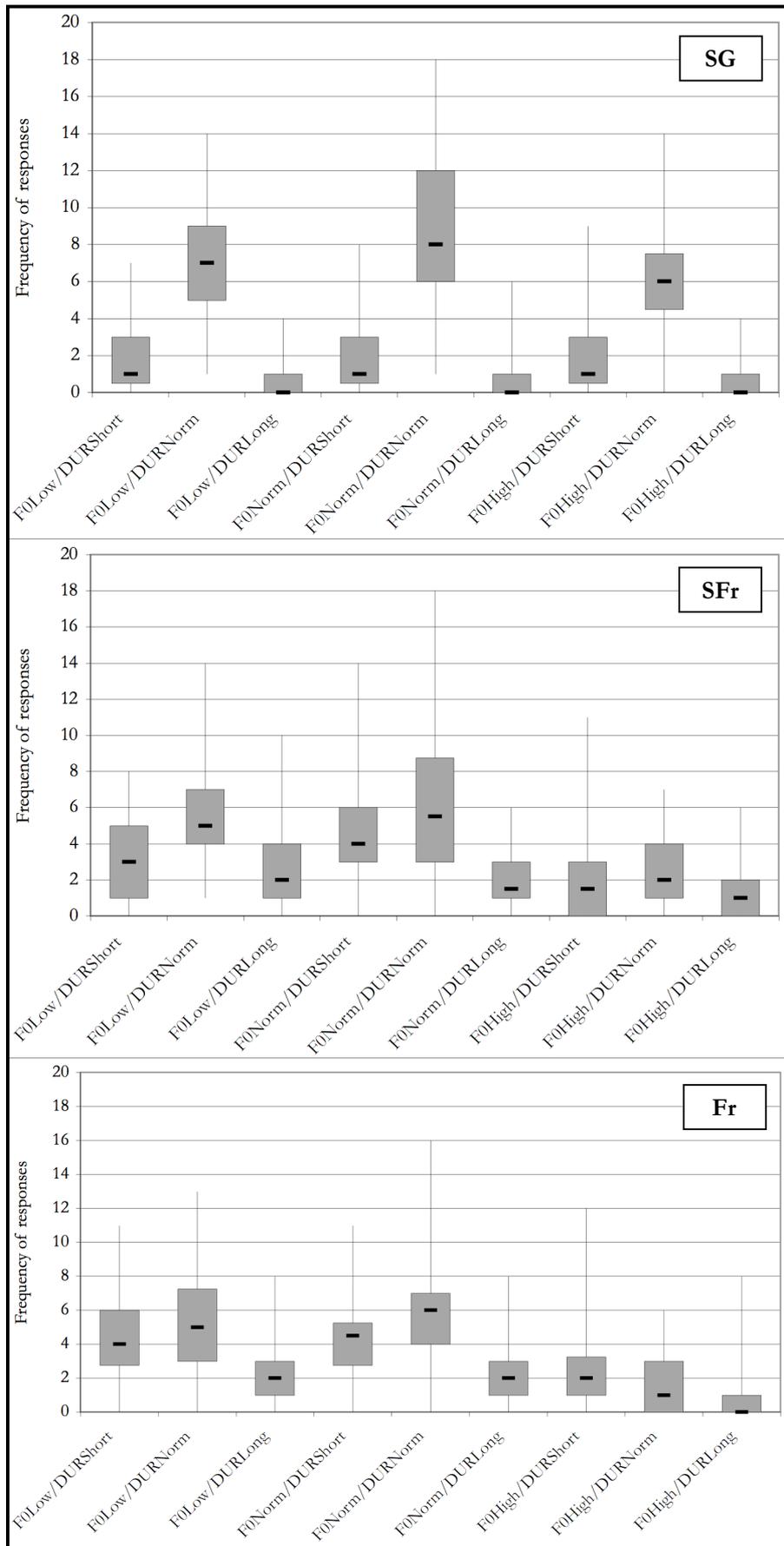


Figure 5-15 – Median (small black rectangles), interquartile range (grey boxes) and minimum/maximum (vertical lines) of response frequency per stimulus type

For all languages, $F0_{Norm}/DUR_{Norm}$ had the highest median frequency and also the largest range. Some subjects preferred it in around two thirds of thirty trials, whereas others (two Frs and one SFr) never perceived it as most rhythmically-natural. One of these Frs chose $F0_{Low}/DUR_{Short}$ in one third of trials, and the other Fr had a more even spread of responses; the SFr chose $F0_{Low}/DUR_{Long}$ in one third of trials. These subjects were not excluded, because their post-test questionnaire suggested that they had understood the task, and some individuals (within any population) could have an opinion on rhythmicality which differs from the average or from the particular speaker who did the original recordings. The minimum/maximum statistics show one or two subjects at either extreme, but the relatively small interquartile ranges (even the largest, SG $F0_{Norm}/DUR_{Norm}$ and SFr $F0_{Norm}/DUR_{Norm}$, fall within four responses of the median) demonstrate that many subjects preferred each stimulus type a similar number of times to many other subjects.

This exploration of intra- and inter-subject consistency has demonstrated the following. SG individuals more consistently than (S)Fr individuals preferred a small range of stimulus types; this intra-subject response variation was not attributable to sentence structure (words/syntax); response patterns differed between subjects, as expected in any population, but a substantial proportion of subjects clustered relatively closely around the average number of preferences for each stimulus type.

5.7.3.3 Post-test questionnaire responses

According to the ratings of task difficulty (1 = very easy, 7 = very difficult) (Figure 5-16), some subjects found it easier than others. Some rated at either extreme, but many rated around 4 to 5. The mean ratings (SG=4.61, SFr=4.59, Fr=4.46) were not significantly different between languages.

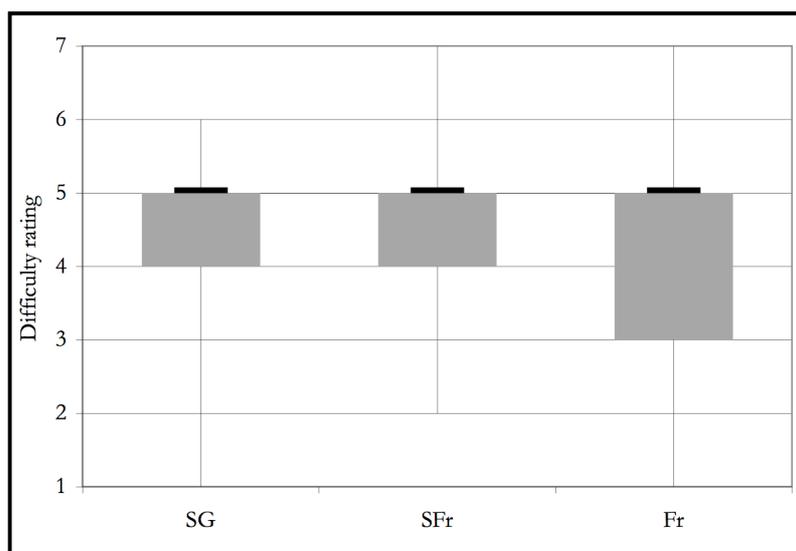


Figure 5-16 – Median (small black rectangles), interquartile range (grey boxes), and minimum/maximum (vertical lines) of difficulty ratings

During the pilot experiment, some subjects made comments like ‘I’m not musical’, which no subjects in the previous experiments (chapters 3-4) had mentioned. In other studies, musically trained participants have performed better than non-musicians in prosodic linguistic tasks including: detecting weak f_0 manipulations in sentences of their native language (Schön et al. 2004) and of a language unknown to them (Marques et al. 2007); identifying affective states cued by prosody in sentences and tone sequences (Thompson et al. 2004); and perceiving lexical stress in another language when their native language lacks this (Kolinsky et al. 2009) (compare with studies on ‘stress deafness’ described in chapter 1, e.g. Dupoux et al. 2008). As Patel and Iversen (2007) highlighted, these studies show correlations and not necessarily causal relationships, and individuals who seek out musical training may have other pre-existing cognitive characteristics which also enhance their processing of linguistic prosody (cf. Kolinsky et al. 2009). Niebuhr (2008, 2009) found that two tasks (both investigating whether utterance-global pitch-based prominence pattern, i.e. rhythm, influences local perceived prominence in German) showed different effects of musical training. The initial task (Niebuhr 2008) required German subjects to consciously listen to prominence patterns as they repeatedly heard a sentence (with various f_0 manipulations) and then selected which syllable they perceived as emphasised. Musicians, but not non-musicians, showed that perceived local prominence depends on utterance-global rhythm. Conversely, musicians *and* non-musicians showed this effect in a subsequent experiment, which made the task less ‘metalinguistic’ (Niebuhr 2009: 103) by including various sentences with lexical-stress-related minimal pairs and different f_0 manipulations, and making subjects ‘shadow’ the sentences. Prominence was judged according to meaning, so explicit attention to acoustic cues was not required, and subjects’ reproductions represented implicit but clear judgments of prominence (Niebuhr 2009).

In the present experiment, the post-test questionnaire asked subjects to detail their musical training, to test whether this correlated with task-difficulty rating. Subjects were divided into four musical training categories: (1) complete lack of training; (2) maximum of five years music tuition (theoretical/practical) in the past, mostly at school; (3) more than five years music tuition, including extra-curricular lessons, stopped only recently or continuing currently; (4) music qualification at ‘Matura’/‘baccalauréat’ (Swiss/French equivalent of British A-Levels) or university level, and/or currently participating in considerable musical activity. Figure 5-17 shows that highly musically trained subjects generally found this linguistic task no easier than those who have had more limited or no formal musical experience. A Spearman Rank correlation of difficulty rating and musical training category fell clearly short of significance ($r_{ho} = -0.035$, $p = 0.686$).

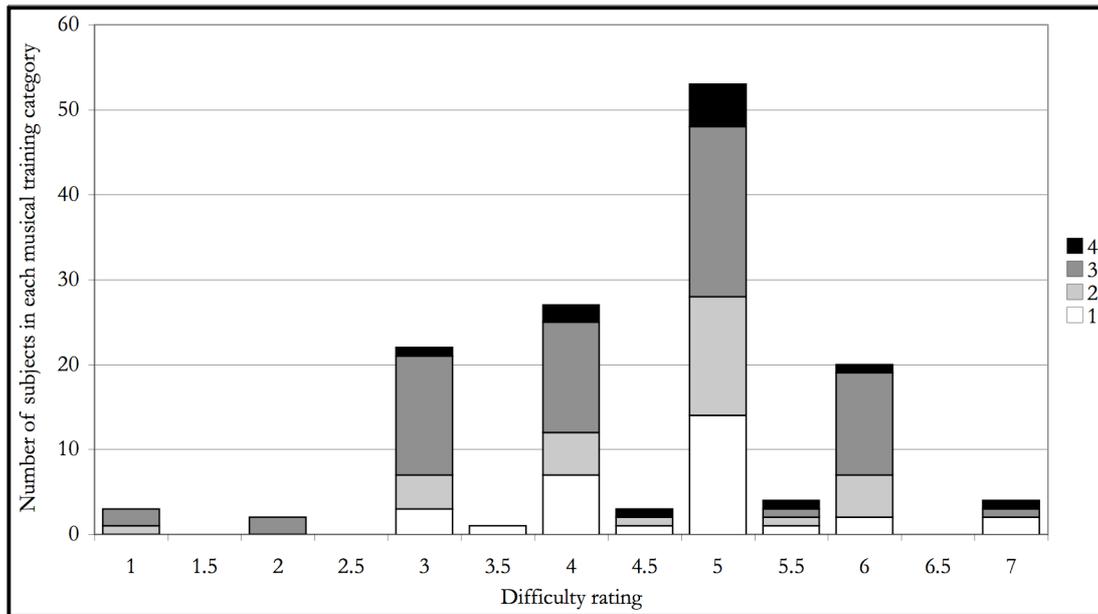


Figure 5-17 – Difficulty ratings split by musical training category (1-4)

Subjects were deliberately given no definition of rhythm, since the pilot experiment showed that naïve subjects needed no explanation of rhythm to complete the task, and it was inappropriate to bias them with a phonetician’s view. Inter-subject variability could have been caused by differing views on what rhythm is, so the post-test questionnaire asked subjects to describe, in their own words, which criteria they used to judge rhythm. We see from Figure 5-18 that these naïve subjects’ self-reported criteria were, considering they received no definition beforehand, remarkably similar to those found in expert definitions of rhythm.

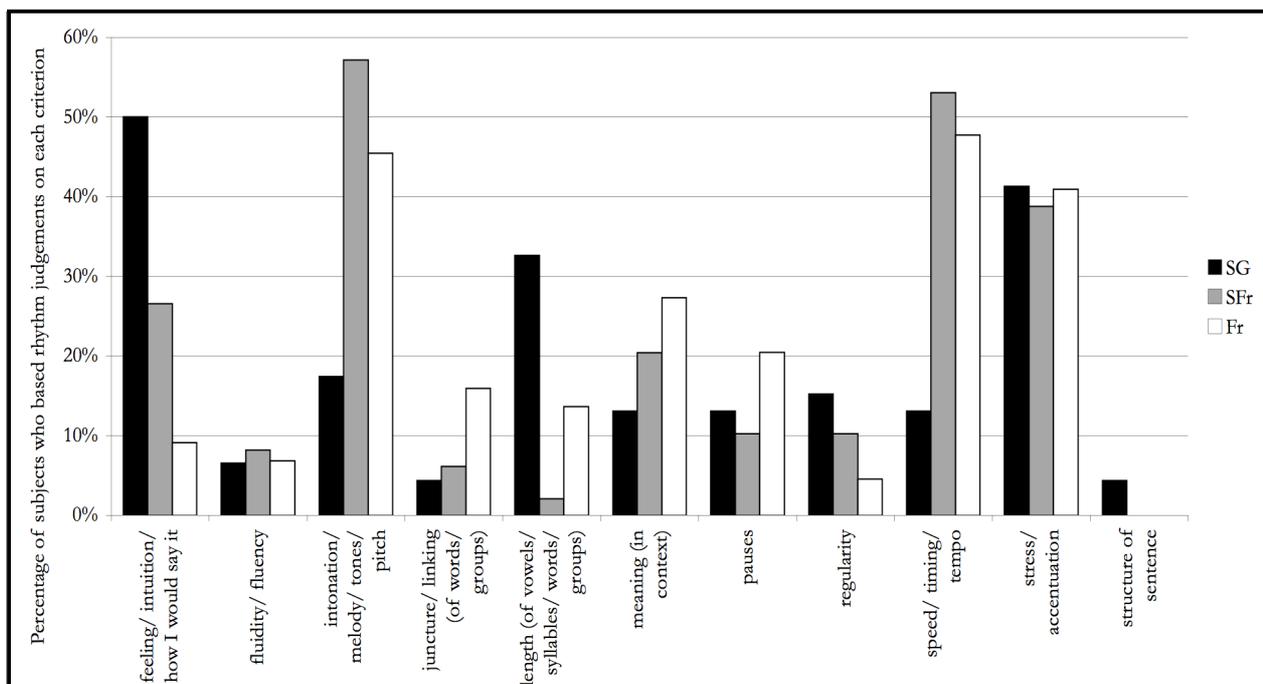


Figure 5-18 – Criteria used by subjects to judge natural-sounding rhythm (English translations – RC); individuals could give multiple criteria

Individuals often based their rhythmicality judgments on several criteria, which illustrates the complexity of rhythm. Many more SGs than (S)FrS said they relied on feeling, intuition or ‘how they personally would say it’; these subjects’ responses were generally no more variable than average, they just could not always put into words the acoustic manipulations to which they attended (some also mentioned more ‘technical’ criteria). Intonation and speed were more important for (S)FrS than SGs, whereas (vowel/syllable) length was more important for SGs than (S)FrS. These self-reports accord with the tent-shaped graphs: tonal and durational manipulations within a certain range mattered to (S)FrS’ perception of rhythmicality; durational manipulations mattered much more than tonal manipulations to SGs’ perception of rhythmicality. Intonation and speed were more important for SFrS than FrS, whereas length, pauses and linking were more important for FrS than SFrS. In all three groups, stress/accenuation was popular.

The post-test questionnaire asked subjects to explain their technique for the task, since this would gain insight on the possible strategies employed in judging a natural-sounding rhythm. Furthermore, inter-subject variability in rhythmicality judgements may have resulted from whether individuals used a systematic or random procedure. Subjects had simply been told the computer program’s practicalities (that a sentence would play with each click on a box, and how to submit their decision). Despite this freedom to adopt their own technique, many subjects proceeded similarly, as Figure 5-19 shows.

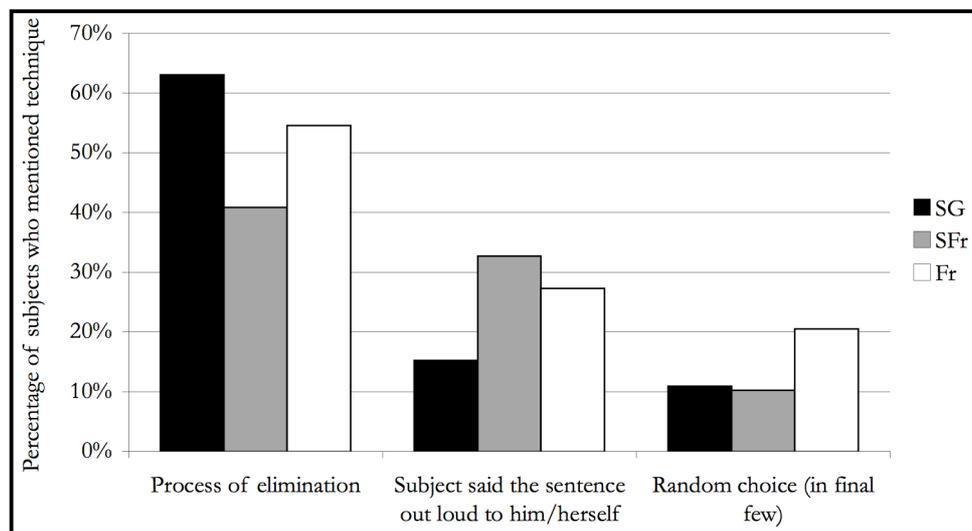


Figure 5-19 – Three most common answers when subjects were asked to explain their technique for the task

A ‘process of elimination’, used by over half the subjects (especially SGs), meant that they clicked through all nine stimuli once, eliminated the worst ones, then went back and compared the best ones (usually three to four) to choose the most rhythmically-natural one. Several subjects, particularly (S)FrS, admitted that they spoke the sentences out loud, simultaneously or successively to the stimuli, hence a transformation from a listening task into an

unrecorded production task too. ‘Random choice (in final few)’ refers to the few subjects who wrote that they completed the task rather randomly, usually relying on intuition, or that their final choice of one amongst two/three ‘best’ ones was sometimes at random.

5.8 Discussion

5.8.1 Interdependence of duration and f0

It was predicted that the duration and f0 excursion of prominent syllables must both be appropriate for a sentence to sound rhythmical, but that subjects’ tolerance level for deviance of each cue might vary between languages, individuals and sentences. If so, subjects would choose the non-deviant $F0_{\text{Norm}}/DUR_{\text{Norm}}$ stimulus as most rhythmical most often, but might sometimes choose other stimuli. This response pattern was indeed observed. Clearly duration and f0 are interdependent in perceived sentence rhythmicality: when the duration or f0 excursion of the medial-prominent syllable was non-deviant, the (non-)deviant state of the other cue contributed significantly to subjects’ rhythmicality judgements. Native language significantly affected the weighting of each cue, since SGs often tolerated tonal deviance but hardly ever tolerated durational deviance, whereas (S)Frs often tolerated some deviance in both duration (shorter) and f0 excursion (lower).

In a similar experiment, Barry et al. (2009) found that non-deviant duration was more important than non-deviant f0 in English and German listeners’ rhythmicality judgements, whereas the cues were weighted almost equally in Bulgarian listeners’ responses, for which Barry et al. (2009) suggested two alternative causes. First, Bulgarian lacks phonological vowel-length contrasts, so these listeners may be less sensitive to alternating long-short manipulations cueing rhythmicality, because their language does not have a durationally trochaic rhythm. Second, the overall Bulgarian responses were influenced by four listeners whose rhythmicality judgements could be seen as idiosyncratic. This inter-subject and cross-linguistic variation demonstrates that different listeners may adopt different strategies for completing the task, which could result from the complexity of rhythm or from the nature of the experimental procedure and stimuli (Barry et al. 2009). The present experiment also provides evidence that cross-linguistic variation in cue weighting could be affected by speakers of different languages perceiving the complex phenomenon of rhythm differently, or by the stimuli and task. These points are now discussed in turn.

5.8.2 Influence of native-language (prosodic) phonology on perceived rhythm

We find evidence in listeners’ independent rhythm definitions that the precise combination of acoustic cues which gives the impression of perceived rhythm is not identical across languages. The (S)Frs’ most popular criteria for judging rhythmicality were intonation (f0-

related), speed (duration-related) and stress/accenuation (potentially both cues). Together these accord with the deduction from (S)Frs' responses to stimuli that the duration and f_0 of the medial-prominent syllable were similarly important for them. The SGs' most popular criteria for judging rhythmicity were intuition, stress/accenuation (potentially both cues) and vowel/syllable length (duration-related). The latter alone accords with the deduction from SGs' responses to stimuli that the duration of the medial-prominent syllable was more important for them than the f_0 .

Subjects' definitions and rhythmicity judgements are explainable with the idea that the properties of their native language could determine which cues they attend to when perceiving sentence rhythm. An appropriate duration contributes more than an appropriate f_0 to SGs' rhythmicity judgements, whose language has complex syllable structure and vowel- and consonant-length contrasts, hence greater variability in syllable duration. Conversely, the cues are weighted more equally in (S)Frs' responses, whose language shows lower variability in syllable duration (cf. for Bulgarian, Barry et al. 2009). Durational cues may be more salient signals of rhythmicity for listeners whose language has markedly contrasting inter-syllable durations. This perhaps explains why durational properties have dominated the impressions of rhythm reported by Germanic-language-speaking linguists, i.e. rhythm typology (see chapter 1).

Tonal properties are equally addressed in this discussion, since f_0 is also important. Rhythm involves prominence, and in this task specifically, the regular pattern of three prominent syllables with rising f_0 and lengthening. Perhaps (S)Frs do not separate prominence and intonation when perceiving rhythm because the relationship between prominence and intonation is closer in Fr compared to that in SG. In Fr, obligatory prominent (lengthened) syllables occur RG-finally, which is also the location of pitch-accents. Some linguists have thus claimed that Fr intonation and prominence are not distinct in form or function (e.g. Rossi 1980); others maintain that prominence and intonation are separate, though closely related (e.g. Di Cristo 1999). Either way, pitch and length could be equally and simultaneously important in these listeners' rhythm judgements. For SG, Fleischer and Schmid (2006) claimed that more pitch-accents seem to occur (in the Zürich dialect) than in standard (northern) German. Häsler et al. (2005) found that in spontaneous and read SG and SstG speech, utterance-global f_0 movements were more negligible than in standard German read speech, and in SG a greater number of IP-internal pitch-accents occurred. In the present experiment, SGs generally did not report (utterance-global) intonation as a rhythmicity-judging criterion, perhaps because this is less salient for them than local prominence-leading f_0 movements. Popular criteria were stress/accenuation and vowel/syllable length, which involve syllable-level f_0 movements and duration. For SGs, the considerable variation in the timing of f_0 rises may make this pitch cue a less useful cue to rhythm, because its varying location relative to syllable boundaries could confuse the perception of regularity. (Across the thirty original sentences, peak-time statistics were: mean=51.43ms after the medial-prominent

syllable's end-point; minimum=-169.08ms (i.e. before the end-point); maximum=228.16ms; standard deviation=96.42ms.) Duration, however, always increased between the start and end of the one prominent syllable. The f0 rise adds to perceptual prominence, as listeners hear the syllable where it starts as prominent, but duration could be the focus of their attention. This would not necessarily be the case for (S)Fr_s, since the rise always occurs within the prominent syllable's boundaries.

In experiment 2 (chapter 4), dynamic pitch and increased length were more effective cues to rhythmic groups for SGs and (S)Fr_s respectively. However, in perceiving a rhythmically-natural sentence, an appropriate length is more important than an appropriate pitch for SGs, and both are similarly important for (S)Fr_s. Experiment 2 required listeners to locate the boundary between two RGs. This experiment involved perceiving as rhythmical a regular prominence pattern across three RGs, in which Fr and SG prominence had to be RG-final and RG-medial respectively; in Fr then, but not SG, boundaries and prominence co-occurred. SG is tightly constrained durationally by vowel- and consonant-length contrasts. Therefore, it is conceivable that SG speakers use and expect tonal (rather than phonologically constrained durational) properties to mark RG boundaries, but cannot tolerate deviance from the constrained length in a pattern of lexically-stressed syllables for a sentence to sound rhythmical. Fr is not durationally constrained by phonological contrasts, so speakers probably have no problem in using and expecting syllable lengthening to mark RG boundaries, which coincide with prominence and pitch movements. For a Fr sentence to sound rhythmical, length and pitch properties of prominent syllables can deviate, but speakers are more tolerant of a decrease in both cues' magnitude, and thus in prominence, than an increase. Their impression of rhythm may less specifically concern an utterance-global prominence pattern across multiple RGs, but rather a more even durational and tonal pattern within RGs, though with boundary marking present. This accords with Vaissière's (1991b: 109) view that Fr rhythm is not like English (or we could infer SG) rhythm in which 'the stream of units of information is thrown into disorder by the intrusion of a recurring strong stress.'

Subtle prosodic differences between SFr and Fr, and northern and southern Fr did not lead to a significant distinction between these subjects' rhythmicality judgements. The SFr recordings rejected for this experiment showed RG-penultimate as well as RG-final f0 rises. Miller (2007) also found that final rises occurred significantly earlier in SFr than Fr, and that SFr had longer phrases with fewer pauses than Fr, which could explain the reported perceived slowness of SFr. If final rising f0 and a faster speed were rhythmically less natural for SFr_s, we might have expected them to prefer F0_{Low} and DUR_{Long} more often than Fr_s. This is evident in the results, but the differences are negligible, which is unsurprising since (according to the post-test questionnaire) most SFr_s (55.1%) knew that the speaker was from France, whilst only 16.3% thought Switzerland, and 10.2% could not say. For Fr_s, 54.5% knew that the speaker was from

France, 4.5% thought Switzerland, and 20.5% could not say, though apparently most of these assumed he was from France but could not pinpoint the region. That SFrS found the task no more difficult than FrS, yet most knew the stimuli were not SFr, and that the two groups' responses did not differ significantly, suggests that negative/positive social attitudes towards one variety or another were not an issue here. The other 18.4% of SFrS and 20.5% of FrS thought that the speaker was from somewhere else, mostly an Anglophone or German-speaking country; several of these commented that his intonation or accentuation/stress sometimes sounded like that of a Germanic language.

5.8.3 Cross-linguistic differences in stimuli

The sentences were designed to be cross-linguistically as equivalent as possible, but inevitable differences occurred concerning prominence position, both its location within RGs, and that the rise often crossed two syllables in SG but not Fr. Another cross-linguistic difference manifested itself in the speed of the recordings. Speakers read the sentences at a comfortable normal pace for them, followed by a slow and a fast repetition, then two further 'normal' repetitions to check that the normal pace was repeatable and contrastable to speed changes. Only normal pace recordings were selected for stimulus creation. The mean length of the thirty selected sentences was significantly different between languages: 1.71secs for Fr (5.32 syllables/sec), 2.13secs for SG (4.25 syllables/sec) [$t(58)=-9.74$, $p<0.0001$]. An obvious explanation is another unavoidable difference in phonological structure. Syllable structure is more complex in SG than Fr, and longer consonant clusters and phonologically long vowels in SG inevitably have greater duration. Alternatively, the Fr and SG speaker could be idiosyncratically fast and slow respectively. The post-test questionnaire asked: 'Did you notice anything unusual about the speaker?' Two SGs thought he spoke slowly, one SFr and one Fr thought he spoke quickly, and one Fr found him slow, which suggests that these speakers' rates were not unusual. Some subjects commented that he spoke clearly, and they alluded that elision, which we would expect in more informal speech with increased coarticulation, was lacking. This seemed more notable in SG than (S)Fr (nine versus six comments), which could be an inevitable consequence of the SG speaker being less used than the Fr speaker to reading aloud sentences written in the language he speaks.

Another consequence of cross-linguistically different syllable structures is that the mean duration of the to-be-manipulated syllable in the original sentences was significantly lower in Fr than SG [$t(58)=-7.296$, $p<0.0001$]. This was one reason for implementing a percentage rather than absolute duration change. The mean f_0 excursion of this prominent syllable was not, however, significantly different between SG and Fr [$t(58)=0.374$, $p>0.05$]. Syllable durations in the centre RG of the original sentences were generally less varied in Fr than in SG (prominent syllable in bold: Fr means = 151ms, 160ms, **189ms**; SG means = 269ms, **309ms**, 200ms). In SG,

the manipulated syllable was surrounded by two generally much shorter syllables in the same RG, whereas in Fr, the manipulated syllable was preceded by a syllable in the same RG and followed by one in the next RG, both of which were generally slightly shorter. It might be argued that the SG stimuli drew more attention to duration than the Fr stimuli did. However, many SGs were clearly aware of f_0 manipulations, given their post-test questionnaire responses. One of their popular criteria for judging rhythmicity was stress/accenuation, which probably involves both cues (though perceptual research is needed), and many commented that the speaker had melodious or monotonous intonation in different sentences. An appropriate duration was just more important for them than f_0 excursion, which could deviate while the sentence remained rhythmically natural-sounding.

The variation between language groups' responses may have resulted from the unavoidable linguistic differences between Fr and SG stimuli. Nevertheless, given that subjects' rhythm definitions reflected their responses to stimuli, it seems unlikely that the cross-linguistic differences in stimuli were the only cause of cross-linguistic variation in responses, and it is also conceivable that the properties of subjects' native language determined which cues they attended to when perceiving sentence rhythm. Moreover, the within-language lexical/syntactic variation between stimuli was found to be unrelated to the overall difference between SG and (S)Fr responses. Rhythm is a complex phenomenon, whereby several potential cues are present, any combination of which a given individual may perceive as more important in defining a natural-sounding rhythm depending on their native language's properties. This is why cross-linguistic research on rhythm perception is needed, even though cross-linguistically equivalent stimuli are hard to design.

5.8.4 Other factors linked to inter-subject variation

It was predicted that subjects' tolerance level for durational and tonal deviance might also vary between individuals and sentences *within* language groups. Indeed within-group inter- and intra-subject variation in responses to stimuli was observed (cf. Barry et al. 2009); this was found to be not attributable to lexical/syntactic differences between sentences. The post-test questionnaire sought information about factors which could cause inter-subject variation: task difficulty, procedure for each trial, and rhythm definition. Some subjects found the task harder than others. Arguably the task requires 'metalinguistic' thought, a conscious awareness of rhythm which is not necessary in everyday speech perception. However, since some reported that they relied on intuition, and since musical training did not correlate with difficulty rating, a conscious awareness of specific acoustic cues or the ability to make an analogy with music-like 'beats' was apparently not essential for subjects to complete the task or find it easy.

Some subjects spontaneously repeated the sentences aloud, which presumably made the task easier for them, perhaps because as they clicked around the randomised matrix of stimuli

they could stop when they perceived the sentence's rhythm to match their own production. In the post-test questionnaire, 22%, 4% and 0% of SGs, SFrs and Frs respectively wrote that they noticed in several trials three stimulus groups within the (randomly presented) nine, or two poles/tendencies in opposite directions away from naturalness, though length/pitch was not always mentioned. Consequently, they might have judged each trial's most natural rhythm whilst keeping in mind a pattern underlying the nine stimuli which they compared across many trials; in which case they might have preferred just two or three stimulus types, hence lower intra-subject variability. The majority, however, apparently dealt with each trial by itself, judging sentences individually without the concept of an underlying pattern, which seems to be a less categorical strategy than recognising a pattern across trials. If the more obviously manipulated stimuli had been used (see §5.4.2.1), more subjects might have recognised a pattern, but those stimuli were not used precisely because they could have been too unnatural.

Given that experts have not agreed on one definition of rhythm, it is unsurprising that not all subjects in each group defined rhythm identically, though, as discussed above, there were strong language-group tendencies. Meaning (in context) was a self-reported criterion for judging a natural rhythm for 13%, 20% and 27% of SGs, SFrs and Frs respectively (cf. Auer and Couper-Kuhlen 1994, Auer et al. 1999; see chapter 1). This clearly demonstrates that they heard the stimuli as meaningful linguistic utterances. Given their comments, some subjects also recognised that a natural rhythm sometimes depends on affective state, speech style and contextual emphasis. This further highlights rhythm's complexity, not just acoustically, but related to higher levels of linguistic structure. For example, if the speaker is conveying some exciting or happy news, the perceived natural rhythm may involve shorter duration (i.e. faster rate) and greater f_0 excursion, rather than longer duration and lower f_0 excursion if the news is boring or sad. Similarly, the perceived natural rhythm may differ between a formal situation or a conversation between friends. Natural-sounding rhythm was judged according to a perceived regular pattern of three prominences, so the appropriate duration and f_0 excursion of the medial-prominent syllable (i.e. the preferred stimulus) could depend on the contextual prominences, i.e. the first and third prominent syllables' durational and tonal properties. Beyond the sentence level, the preferred stimulus could depend on the position of each sentence in the context of a longer discourse, which listeners were free to imagine. The surrounding sentences could affect whether the medial RG would be more/less emphasised than the other RGs in conveying the sentence's intended meaning, and thus whether the utterance-medial prominence should be more/less acoustically salient than surrounding prominences, for a perceived natural rhythm. Although these sentence-local and wider contextual factors may influence perceived rhythmicity, responses were not significantly affected by between-sentence lexical/syntactic variation here.

5.9 Conclusion

First, this experiment has demonstrated a viable method for tapping into naïve listeners' intuitions about speech rhythm. They were not given a definition of rhythm or told how to search for the most rhythmically-natural sentence, yet they could complete the task systematically. Moreover, they identified what rhythm meant for them, which included not just durational properties like timing and length, but tonal properties including accentuation and intonation. This alone certainly encourages the idea that rhythm research should investigate acoustic cues other than duration.

The findings of experiments 1 and 2 (chapters 3-4) suggested that investigating durational-tonal interdependence in rhythm perception is worthwhile; they also demonstrated that the perceptual interdependence of duration and f_0 differs between linguistically meaningful and meaningless sounds, and between varied and identical syllable sequences which are more and less frequent in natural language respectively. Therefore, as Niebuhr (2009) also argued, rhythm perception experiments must present stimuli which listeners hear as natural meaningful utterances, to mimic the communication process and make the results generalisable to real speech perception. Previous rhythm perception studies which presented speech in unfamiliar languages (e.g. Miller 1984) or segmentally degraded stimuli (e.g. Ramus et al. 2003, White et al. 2007) did not reveal how listeners perceive rhythm in their native language, the linguistic structure and properties of which they know and may be influenced by when perceiving rhythm. Other studies used linguistically meaningful stimuli, but acoustically concerned only duration (timing) (e.g. Donovan and Darwin 1979, Magne et al. 2004, Scott et al. 1985), or only f_0 (Niebuhr 2009). Barry et al. (2009) investigated the interdependence of prominence cues, but their stimuli were meaningless, thus losing the concern for applicability to real communication.

The present experiment was apparently the first in which naïve listeners directly judged the rhythmicality of sentences involving more than simply durational manipulations. The results demonstrate that for a sentence to sound rhythmically-natural, the duration and f_0 excursion of prominent syllables must both be appropriate within a certain range, which depends on native language and potentially other factors. These results are more generalisable to speech perception than those from nonsense stimuli or prosodically stylised stimuli. The stimuli were resynthesised from naturally produced utterances, which allowed tight control over the manipulation made, whilst preventing the stimuli from sounding artificial. Only two subjects thought it might have been a computerised voice, and several thought it was one very gifted speaker producing such finely different sentences! Nevertheless, these were read sentences with a controlled prosodic structure. A further investigation could use this experiment's method with spontaneous speech stimuli (as in e.g. Allen's (1972) beat-tapping experiment), to see if listeners' rhythmicality judgements are still as systematic. It would also be interesting to investigate a domain longer than

the sentence, to examine further the effect of discourse context. However, if spontaneous speech were used, it would be difficult to assess the cross-linguistic equivalence of stimuli. When stimuli involve linguistic structure, this naturally varies between languages, and the more linguistically complex the stimuli, the harder they are to control for experiments. In this experiment, the inevitable differences between SG and Fr stimuli could partly have caused the different response patterns, though this is unlikely to be purely a product of the experiment. It is probable that native speakers of different languages could perceive the complex phenomenon of rhythm in different ways, depending on which cue(s) out of the many available in the complex signal is/are most effective given their language's prosody. A testable prediction is that if listeners do this experiment in a language unknown to them, their preferences for stimuli should be different from the preferences of native speakers of that language; the non-native listeners would be influenced by their own native prosody, and so would attach different importance to the various rhythmicity cues compared to native listeners. Different cues might also be more useful in various listening conditions, e.g. noisy versus quiet situations.

Two points are now overwhelmingly clear. First, rhythm research should investigate both f_0 and duration, because these are interdependent in rhythm perception. Second, investigating rhythm perception from a cross-linguistic perspective is essential, if we desire a universal view of the phenomenon, not biased by the weighting of cues in any particular language. The following chapter reports an experiment which takes a first step in applying the findings from this chapter's perceptual task to the sort of production investigation that has recently become popular in rhythm research.

PVIs which account for the acoustic multi-dimensionality and language-specificity of perceived rhythm

6.1 Summary

The previous experiments concerned perceptual tasks, since rhythm is defined in this thesis as a perceptual phenomenon. These demonstrated that f_0 and duration are interdependent in the perception of isolated syllables, rhythmic groups and sentence rhythmicity, and that the relative weighting of tonal and durational cues depends on listeners' native language. The experiment reported here investigates whether this perceptual finding can be applied to production data, to make a Pairwise Variability Index (PVI) perceptually informed. The relative weighting that an appropriate duration and f_0 contributed to listeners' rhythmicity judgements is calculated, and these language-specific weighting values are incorporated into combined durational-tonal PVIs, to quantify rhythm in SG, SFr and Fr. The results demonstrate that produced rhythm, when quantified in a way that accounts for the acoustic multi-dimensionality and language-specificity of perceived rhythm, is less cross-linguistically divergent than rhythm quantified with language-universal durational metrics.

6.2 Rhythm metrics

6.2.1 Problems

Several 'rhythm metrics' have been proposed to quantify rhythm which calculate the durational variability of phonetic/phonological units, resulting in a summary index, a discrete value to represent rhythm and compare it cross-linguistically. Some calculate variability utterance-globally: ΔV , ΔC (the standard deviation of vocalic- or consonantal-interval duration); VarcoV, VarcoC (the sd divided by the mean vocalic- or consonantal-interval duration, multiplied by 100). Alternatively, Pairwise Variability Indices (PVIs) calculate the difference (usually durational) between the members of each successive pair of units (usually vocalic or consonantal intervals), then all pairwise differences are added up and their mean taken. (For an overview of the various metrics, including methodological information, see Widget et al. 2010 and Fletcher 2010.) Recently rhythm metrics have come under attack (see Arvaniti 2009, Barry et al. 2009, Kohler 2009a). Here only two problems are considered.

The first problem is that rhythm metrics have become predominantly durational, except a few studies that have calculated non-durational PVIs (e.g. Ferragne 2008: intensity, Low 1994: amplitude, spectral dispersion). Several studies concerned with (categorical/continuous) rhythm 'types' have plotted durational rhythm-metric values on a two-dimensional graph, e.g. Grabe and Low (2002) (x-axis, vocalic PVI; y-axis, consonantal PVI), Nolan and Asu (2009) (x-axis, syllable PVI; y-axis, foot PVI). However, f_0 and duration are interdependent in rhythm perception

(according to chapter 5). A potential solution could be to calculate ‘tonal’ metrics using f_0 measurements which could complement durational metrics in cross-linguistic comparisons of rhythm. Rhythmically distinct languages, which are not differentiated on a typical durational rhythm-metric plot, might be distinguished if their tonal PVI were plotted against their durational PVI. However, adding yet another dimension to the already overcrowded discussion of rhythm typology is unappealing given that this thesis aims to provide fresh perspectives, which could get round the impasse that rhythm research has arrived at after a prolonged focus on rhythm typology.

The second problem is that rhythm metrics only capture the speech signal’s physical nature, and do not reflect the relative significance of each acoustic cue which differs cross-linguistically in rhythm perception (according to chapter 5). For example, let a durational PVI of 68 and 42 represent the rhythm of languages A and B respectively. It is known that for speakers of A, durational variability creates a strong impression of perceived rhythmicity, but for speakers of B, tonal variability creates a stronger impression than durational variability. Thus the acoustic durational variability of ‘42’ is of little consequence to speakers of B, but that of ‘68’ is highly significant to speakers of A, when they all perceive rhythm in their native language. If the duration-based metric is applied universally, the resulting figures (68, 42) do not have any meaningful interpretation in terms of cross-linguistic differences in perceived rhythm.

This point was repeatedly made by Barry et al. (2009: 78), who stated that empirically grounded conclusions concerning the link between produced and perceived rhythm are rare in rhythm-metric studies, except those in which listeners discriminated, from segmentally degraded speech stimuli, languages that have different rhythms when quantified with metrics (e.g. Ramus et al. 2003, White et al. 2007). From their results, White et al. (2007: 1012) argued that listeners attended to and interpreted the durational variability represented by the rhythm-metric scores calculated from the stimuli, which shows ‘strong support for the use of rhythm metrics in classifying perceptually salient aspects of speech rhythm.’ Barry et al. (2009: 79) explicitly wanted to address ‘the relationship between the concrete rhythmic measure and the rhythmic impression of the utterance that produced it.’ They calculated durational PVIs for poetical verses with various metrical patterns spoken in three languages, and found that similar PVI values represented different perceived rhythms, and divergent PVI values represented similar perceived rhythms. (Their subsequent perceptual experiment, which showed that non-durational cues were significant in rhythm perception, was reported in chapter 5.) Given their results, Barry et al. (2009) criticised PVIs and implied scepticism towards the metrics’ future usefulness.

6.2.2 Possible solution

Ultimately we may question whether applying durational metrics language-universally is justifiable, if they cannot capture the acoustic complexity and language-specific nature of

perceived rhythm. Nevertheless, one possible solution, proposed in the following experiment, is to combine duration and f_0 in a PVI, and weight each cue's contribution according to its significance for perceived rhythm in each language investigated. This makes the PVI perceptually informed. The PVI was chosen here over other metrics because it has been popular in terms of the number of studies featuring it (Nolan and Asu 2009), and its normalisation procedure and regard for successive units are an advantage over utterance-global metrics like ΔV and ΔC (for evaluation of various metrics see e.g. Arvaniti 2009, Barry et al. 2003, Ramus 2002, White and Mattys 2007a).

Most PVI studies have measured phonetically defined (vocalic/consonantal) intervals, perhaps influenced by the reasoning in two influential papers: Ramus et al. (1999) developed rhythm metrics from the idea that infants perceiving rhythm have no phonological knowledge so rely on salient acoustic properties of the signal; Grabe and Low (2002), in a cross-linguistic PVI study, wanted to avoid subjective segmentation decisions based on phonological criteria for the languages in the sample which they did not speak. Conversely, Nolan and Asu (2009) calculated syllable and foot PVIs, reasoning that 'if we assume that languages do have rhythm, it is surely reasonable to suppose that this is a property which can be informed by the phonological structure, part of which [...] is concerned with grouping smaller elements into larger units [e.g. syllables and feet].' (p.69) Nolan and Asu (2009) admitted that the syllable is often difficult to determine (perhaps a reason for avoidance in PVI studies), but that it is central in phonological structure, which adult native speakers, whose rhythm is often under investigation, have learned.

In the present experiment, eight PVIs are calculated (per language): durational, tonal, combined, weighted and each of these for vocalic intervals and syllables (to compare phonetic and phonological approaches). Combined PVIs incorporate durational and tonal variability, as do weighted PVIs, but these include language-specific weighting for each cue (see §6.4.3.2). (Since f_0 is lost during voiceless consonants, particularly frequent in SG, there is little point in calculating non-durational consonantal PVIs.)

6.3 Hypothesis

6.3.1 Durational PVIs

Previous studies have reported a range of durational PVI values for various languages. It is predicted that vocalic and syllable durational PVIs will be relatively high within this range for SG and relatively low for (S)Fr (see chapter 2 and below for reasons). Evidence comes from Galloway (2007) and Schmid (2001) for SG, Galloway (2007) for SFr, and Grabe and Low (2002), Lee and Todd (2004) and White and Mattys (2007a) for Fr.

6.3.2 Tonal PVIs

No previous studies have calculated tonal PVIs, so the predictions are based on general observations of each language's prosody (see chapter 2 for references). It is predicted that SG and (S)Fr will have tonal PVIs similar to their durational PVIs, because increased duration and substantial f_0 movement co-occur in both languages. In SG, prominent syllables have phonologically long or short vowels, complex syllable structure, and most often a substantial f_0 rise; these prominences contrast with non-prominent syllables, which do not receive pitch-accents, and have reduced vowels and often smaller consonant clusters. Therefore, durational and tonal variability is high. Fr has no phonological length contrasts or vowel reduction, and less complex syllable structure, hence the lower durational variability. Yet its durational (and indeed tonal) PVIs are not (even) lower because rhythmic-group-final (prominent) syllables are lengthened and have substantial f_0 movements.

However, other factors might affect the tonal PVI values, particularly since this metric measures successive variability. We might observe lower tonal (than durational) PVIs for SG, because f_0 rises often begin late in the prominent syllable and continue into the next (see chapters 2 and 5). Thus adjacent syllables 'share' the f_0 excursion, each having a smaller f_0 movement than a one-syllable rise, so they are less varied tonally than durationally. Conversely, we might observe higher tonal (than durational) PVIs for Fr, because if speakers frequently produce optional group-initial prominent syllables, which have substantial f_0 movements but no lengthening (see chapter 2), several adjacent syllables would be more varied tonally than durationally. For SFr, Miller (2007) reported that group-final rises started earlier and group-initial rises started later than in Fr, bringing the two rises closer. Tonal PVIs might be higher for SFr than Fr, since large f_0 movements might be more evenly spaced in SFr compared to longer stretches of non-prominent syllables in Fr.

6.3.3 Weighted PVIs

No previous studies have calculated weighted PVIs. Experiment 3 (chapter 5) found that an appropriate duration of prominent syllables contributed more to SG listeners' rhythmicity judgements than an appropriate f_0 did, whereas an appropriate duration and f_0 contributed to (S)Fr listeners' judgements to similar extents. Therefore, it is predicted that in SG, duration will have greater weighting than f_0 , whereas in (S)Fr, the cues will have more equal weighting. If the durational and tonal PVIs for each language emerge as similar, the weighted PVIs will be similar to these. If the tonal PVIs emerge as lower and higher than durational PVIs in SG and (S)Fr respectively, the SG weighted PVIs should be similar to SG durational PVIs, whereas the (S)Fr weighted PVIs should be somewhere between (S)Fr durational and tonal PVIs.

The point of the experiment is to suggest a method for relating perceived rhythm to a quantification of produced rhythm, and **not** to generate new figures that will better categorise

these languages into rhythmic types. Even if the weighted PVIs emerge as similar to the durational ones, at least they include an element that gives the values meaning in perceptual terms.

6.4 Method

6.4.1 Reading text

Most rhythm-metric studies have elicited speech by having speakers read carefully constructed sentences (e.g. Ramus et al. 1999, White and Mattys 2007a) or a longer text, often ‘The North Wind and the Sun’ in translation (e.g. Grabe and Low 2002). This text was read by the present experiment’s subjects in their native language (SG ‘De Biiswind und d Sune’; Fr ‘La bise et le soleil’; appendix 8.4.1). Therefore, the data were comparable between languages and with other studies which used this text. Generally in phonetics, there is debate over the usefulness of speech data elicited from reading (see e.g. Kohler 2000). For rhythm research, Arvaniti (2009) questioned the use of read speech, because rhythm-metric values for spontaneous speech, unlike those for read speech, did not generally separate languages into rhythm types; this may have resulted from intra- or inter-speaker variation (see also e.g. Barry et al. 2003, Grabe 2002). It is irrelevant to the present experiment whether the elicitation method gives values which distinguish rhythm types. No claim is made that the findings are generalisable to spontaneous speech; that could be explored in future experiments if weighted PVIs turn out to be useful.

Moreover, it was interesting to observe whether rhythm in SG and SFr read speech differed from that previously observed in unscripted speech. SG oral dialects are being increasingly written in text-message and Internet communication. Many SGs commented that reading the text was easy, since they read and write SG daily in text messages and online social networking. This experiment investigated whether their rhythm was unusual when reading. For Fr, Simon (2003: 5) claimed that many studies ‘showed that from read sentences it was difficult, if not impossible, to identify regional prosodic characteristics, probably due to the bias introduced by the “reading prosody”, which seems very standardised for all French speakers’ (translation RC). However, Miller (2007: 147) found that although SFr prosody sounded closer to Fr in read speech than informal-interview speech, the read speech did reveal ‘important small-scale differences’ between SFr and Fr in phrasal organisation and intonation. The present experiment further investigated this difference between SFr and Fr in the prosody (specifically rhythm) of read speech.

6.4.2 Subjects and procedure

Most subjects in experiment 3 (chapter 5) agreed to be recorded whilst reading the text, after the listening task, and they were still naïve to the purpose. Ten subjects’ speech data per language were selected for acoustic analysis (Table 6-1).

Language	Total	Age (years)		Sex		Region
SG	10	Range	20–37	Male	5	From Zürich city; speak Zürich German
		Mean	26.2	Female	5	
SFr	10	Range	19–27	Male	5	From the Neuchâtel canton; most from Neuchâtel town, a few from La Chaux-de-Fonds (twelve miles away)
		Mean	21.6	Female	5	
Fr	10	Range	18–38	Male	5	From greater Paris
		Mean	23.9	Female	5	

Table 6-1 – Summary of subjects

SGs and Frs were recorded in the sound-attenuated studio in Zürich University Phonetics Laboratory and Cambridge University Phonetics Laboratory respectively. SFrS were recorded in a quiet room in Neuchâtel University. In Cambridge the equipment was a *Marantz* PMD670 solid-state recorder, and a low-noise condenser *Sennheiser* MKH40P48 microphone with a cardioid frequency response. In Switzerland the equipment was a portable *Nagra Ares-M II* recorder with an attached mono omnidirectional microphone. For all recordings, the mode was 16 bit linear PCM, with a 44.1 kHz sample rate, and the file was saved as .wav format, then transferred onto a *MacBook* (Mac OS X.5) via a USB cable.

Before the recordings, subjects were given time to read the text. Several SGs made notes on the text, mainly vowel graphemes, because even within one canton subtle dialect differences exist, mainly in vowel quality, and because there is no ‘standard’ SG orthography; although there must be a certain (perhaps subconscious) consensus for communication, each individual ultimately can spell according to what makes sense personally. For the recordings, all subjects were instructed to read the text four times, speaking at a rate and in a style which was normal and comfortable for them, and resting between each repetition. It was thought that by the final time the speaker would read it fluently, but in fact most speakers read the last three fluently. Each subject’s final repetition was selected for analysis, unless it contained disfluencies, in which case the most fluent of the previous ones was selected.

6.4.3 Analysis

6.4.3.1 Acoustic measurements

The selected repetition of each subject’s recording was opened in *Praat* (Boersma and Weenik 2008-2009; version 5.1.04) and displayed with waveform and wideband spectrogram. The *Textgrid* function was used to mark start- and end-points of intervals (vowels and syllables). Since speech is a continuous flow of information, the task of segmenting it is inherently difficult and paradoxical. The criteria used for vowel-consonant segmentation were those of Peterson and Lehiste (1960), a few points of which are noteworthy. All start- and end-points of intervals were

located at the start or end of a glottal period and at the zero-crossing on the waveform. Vowels followed or preceded by plosives, nasals, fricatives or laterals were generally unproblematic, as discrete landmarks occurred in the signal, for example: the abrupt offset of vocalic formant structure at a plosive closure; the onset of high energy frication at a vowel-fricative boundary; an abrupt change in the waveform shape and the onset (offset) of a nasal formant structure, particularly a low F1, at a vowel-nasal (nasal-vowel) boundary; the onset of periodic striation at F1 after a plosive burst or fricative; the spectrographic transient at a lateral release into a vowel. These examples cover the SG and most of the Fr text. The Fr approximants /ɥ j w/ were more problematic. Phonologists regard them as consonants (e.g. Tranel 1987); phonetically, a smooth transition of formants usually occurs from approximant to vowel and vice versa. Juncture was determined by examining the waveform amplitude, spectrogram intensity and formant structure. The point where a fairly sudden change in amplitude/intensity occurred during the formant transition was marked; sometimes auditory perception also played a role. The Fr text contained two adjacent vowels, which were mostly marked as one interval, since speakers produced a diphthong-like rapid transition. In some speakers, a clear separation was evident with glottalised periods; these vowels were marked as separate, since they were more likely to be rhythmically relevant as perceived separate units.

Syllabic segmentation requires phonological and phonetic consideration. A phonetic and phonological syllable have been differentiated in discussions over the concept (e.g. Blevins 1995, Fudge 1969). As noted above, few rhythm-metric studies have measured syllables. In calculating syllable PVI, Deterding (2001) took account of both phonological rules and phonetic realisations in his syllabification method for British and Singapore English. Nolan and Asu (2009) used Deterding's (2001) strategy for English, and they followed accepted phonological rules for Estonian and Spanish syllabification which they found less controversial than in English. Likewise the present experiment considered accepted syllabification rules from phonological descriptions and speakers' variable phonetic realisations. Phonologists maintain that (S)Fr prefers open syllables (i.e. syllable breaks generally occur post-vocally/pre-consonantly) and that resyllabification also occurs across some word boundaries to maintain a (roughly) CV.CV.CV.... structure (e.g. Walker 2001: 27, Post 2000: 97; for intricacies and exceptions to these generalisations, see Walker 2001). Theoretically, (S)Fr does not generally have consonant clusters larger than two; optional schwas are realised and epenthetic schwas are inserted to maintain this situation (Walker 2001). However, the (phonetically variable) pronunciation or suppression of schwas depends on many interacting phonological, morpho-syntactic and stylistic factors, e.g. in informal and faster speech schwas are pronounced less often, sometimes resulting in larger consonant clusters (Walker 2001). The present experiment syllabified according to what was produced, e.g. different speakers pronounced *arriverait* as [a.ʁi.və.ʁɛ] or [a.ʁi.vʁɛ]. For SG,

syllabification (following Fleischer and Schmid 2006, Reese 2007) was straightforward. Like in (S)Fr, across some word boundaries in SG, a CVC(C) syllable followed by a VC syllable is resyllabified into CV.C(C)VC. Deterding's (2001) reason for measuring syllable (not vowel) duration was that many syllables in the conversational British English sample lacked vowels, i.e. had syllabic consonants. In the present experiment, where a syllabic [l] occurred (some SGs occasionally produced this rather than [əl] in 'Mantel') it was not treated as a separate syllable, because f_0 movement was only calculated during vowels, so this syllable would have had no vocalic f_0 value. The [l] was syllabified with the preceding [t] and following vowel (always [ɒ] or [æ]), e.g. [mɒn.tlɒb̥.tso.ɡə].

Three *Praat* scripts were written for extracting vowel durations (ms), syllable durations (ms), and vowel f_0 excursions (semitones). Syllable f_0 excursions were not calculated, because perceptually most relevant f_0 movements occur during steady-state vowels rather than consonants (House 1990), and f_0 would be lost during many consonants due to voicelessness and transitory perturbations. For each subject, all data extracted by the scripts were transferred to *Excel*. The script for f_0 extracted the minimum and maximum f_0 within each vowel, then these values were checked manually by visual inspection of the spreadsheet and spectrogram, and some discrepancies due to non-modal voice were corrected. Occasionally vowels were so laryngealised that *Praat* was unable to compute f_0 ; these were excluded from analysis.

Excursion was calculated as the absolute difference between the minimum and maximum f_0 , i.e. rises and falls were not differentiated. If the negative excursion value for falls had been retained, this would have had odd consequences, since the PVI deals with absolute differences. For example, let three successive vowels have excursions of 1.5, -1.8 and 3.2 st respectively (rise-fall-rise). For V1 and V2, the absolute difference between them is 3.3, their absolute mean is 0.15, and this difference divided by this mean is 22. For V2 and V3, the absolute difference between them is 5, their absolute mean is 0.7, and this difference divided by this mean is 7.14. Therefore, measured by a normalised PVI, the variability between V1 and V2 (22) emerges as higher than that between V2 and V3 (7.14), which is clearly not true. A raw (non-normalised) PVI, which calculates differences rather than differences divided by means, could better represent variability, but this possibility was rejected, as inter-speaker variation in pitch range could result in wide-ranging PVI values. Another reason to treat f_0 excursion as movement regardless of direction is the lack of data bearing on whether rises are perceptually more/less prominent than adjacent equally sized falls. In experiment 1 (chapter 3), falling vowels were perceived as longer than level vowels more often than rising vowels were, so perhaps falls were more prominent than rises. Yet this tells nothing of the perceptual prominence relationship between adjacent rises and falls in continuous speech. Two further justifications for stripping falls of their negative sign came from the present experiment's recordings. First, the majority of

substantial f0 excursions were rises, and falls were generally microfluctuations (apart from IP-finally) so they hardly affected variability anyway. Second, syllables sounded prominent if their f0 excursion was greater than in surrounding near-level syllables, whilst their direction of f0 movement seemed less important for perceived prominence.

6.4.3.2 Weighting values

This section only concerns data from the perceptual experiment in chapter 5, and not the recordings for this PVI experiment. In experiment 3 (chapter 5), listeners judged which stimulus of nine had the most natural rhythm. The nine stimuli were lexically identical, but the medial-prominent syllable in each was acoustically manipulated: 3 duration conditions (DUR_{Short}, DUR_{Norm}, DUR_{Long}) co-varied with 3 f0 excursion conditions (F0_{Low}, F0_{Norm}, F0_{High}) ('Norm' means not deviant from the original recording). From these data we can find out, for each language, the relative weight that the variables 'non-deviant duration' and 'non-deviant f0 excursion' (of stimuli) contributed to whether a stimulus was perceived as most rhythmically-natural. To do this, three separate logistic regression analyses (generalised linear mixed models) were run, one on each language-group's perceptual task responses. According to Field (2005: 220), the binomial logistic regression equation for two variables is:

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_1 + b_2 X_2 + \varepsilon_i)}}$$

In our case:

- $P(Y)$ is the probability that a sentence is judged the most rhythmically-natural;
- e is the base of natural logarithms;
- b_0 is a constant;
- ε is a residual term;
- X_1 and X_2 are the predictor variables 'non-deviant duration' and 'non-deviant f0 excursion', which could be relabelled X_{dur} and X_{f0} ;
- b_1 and b_2 are coefficients attached to the respective predictor variables; these b-coefficients (b_{dur} , b_{f0}), which the logistic regression modelling process estimates based on input data (see below), indicate the weight of each predictor variable's contribution to the probability that a sentence is judged most rhythmically-natural.

The data input to the model were the dependent variable (Y) *response* [1 = the stimulus judged the most rhythmically-natural; 0 = any of the remaining eight out of nine stimuli not judged the most rhythmically-natural], and the predictor variables (X_{dur} , X_{f0}) as follows.

X_{dur} : the difference in duration (ms) between the to-be-manipulated syllable in the recorded sentence and this manipulated syllable in the stimulus. DUR_{Norm} (non-deviant) = 0ms; DUR_{Short}, DUR_{Long} (deviant) = $\pm 35\%$ ms decrease/increase on the original syllable duration (e.g. Table 6-2).

	Example 1 (ms)	Example 2 (ms)
Original syllable duration	100	200
DUR _{Norm}	100	200
DUR _{Short}	65	130
DUR _{Long}	135	270
X_{dur} for DUR _{Norm}	0	0
X_{dur} for DUR _{Short}	-35	-70
X_{dur} for DUR _{Long}	35	70

Table 6-2 – Examples of X_{dur} values depending on original syllable duration

X_{f0} : the difference in f0 excursion (semitones) between the to-be-manipulated syllable in the recorded sentence and this manipulated syllable in the stimulus. F0_{Norm} (non-deviant) = 0st; F0_{Low}, F0_{High} (deviant) = ± 3 st decrease/increase on the original syllable excursion (e.g. Table 6-3).

	Example 1 (st)	Example 2 (st)
Original syllable excursion	4	7
F0 _{Norm}	4	7
F0 _{Low}	1	4
F0 _{High}	7	10
X_{f0} for F0 _{Norm}	0	0
X_{f0} for F0 _{Low}	-3	-3
X_{f0} for F0 _{High}	3	3

Table 6-3 – Examples of X_{f0} values depending on original syllable excursion

The ‘lmer’ function within the R software environment was used¹. The random effect *subject*, which introduced adjustments to the intercept grouped by each subject, was necessary because each subject responded to several trials (Baayen 2008). A non-stepwise forced-entry method was appropriate, as there was no reason *a priori* to enter the predictors (X_{dur} , X_{f0}) in a particular order. Table 6-4 shows the b-coefficients outputted from each language’s model, and the X -standardised b-coefficients calculated from these². Non-standardised b-coefficients

¹ The formula was:

```
> [SG/SFr/Fr]expt4.lmer = lmer(response ~ Xdur + Xf0 + (1|subject), data = [SG/SFr/Fr]expt4,
family = "binomial")
```

² X -standardisation is equivalent to standardising the duration and f0 excursion data (X -variables) before input to the model. Indeed b-coefficients identical to those in the right-hand column of Table 6-4 were obtained by running the same three regression models with pre-standardised data. According to Menard

indicate how much a 1ms or 1st deviation contributes to the judgement of whether the stimulus sounds most rhythmically-natural. However, a 1ms difference between stimuli is less perceptible than a 1st difference. Therefore, standardisation is necessary to find the contribution of each variable relative to the other, regardless of whether duration happened to be measured in ms (or cs, secs etc.) and f0 in st (or Hz, ERB-rate etc.).

Language	Variable (X)	b-coefficient (b)	Standard deviation of X (sdX)	X-standardised b-coefficient (sdX × b)
		** p<0.0001 * p=0.59		
SG	dur	-0.002**	91.044ms	-0.150
	f0	-0.006*	2.450st	-0.015
SFr	dur	-0.004**	55.566ms	-0.231
	f0	-0.109**	2.450st	-0.268
Fr	dur	-0.006**	55.566ms	-0.346
	f0	-0.124**	2.450st	-0.304

Table 6-4 – Output from each regression model

To make sense of what these numbers mean, we can recall the logistic regression equation, which predicts the probability that the dependent variable occurs (i.e. the sentence is judged to have the most natural rhythm)³:

$$P(\text{sentence has most natural rhythm}) = \frac{1}{1 + e^{-(b_0 + b_{dur}X_{dur} + b_{f0}X_{f0} + Z_i + \epsilon_i)}}$$

(2004), various approaches have been suggested for obtaining standardised logistic b-coefficients, some partially (or X-) standardised (i.e. only incorporating variance in the X-variables), and some fully standardised (i.e. also incorporating variance in the dependent (Y) variable). Fully standardised b-coefficients are needed to compare between models the effects of X-variables on the Y-variable, because we cannot assume that the Y-variable in every model has the same variance (Menard 2004). However, the standard deviation of Y can only be estimated indirectly, because during logistic modelling the dichotomous responses (0,1) are transformed to logarithmic odds values (see Menard 2004). Since the point of standardisation here is to compare the relative weighting of duration and f0 excursion (X-variables) *within* each language (i.e. *not* between models), full standardisation, only an estimate, was not implemented.

³ The reader may notice that this equation includes Z, which was not in the previously given equation (§6.4.3.2). This indicates that the model, from which the b-coefficients were estimated, accounted for the random effect of *subject* (see e.g. Baayen et al. 2008, Faraway 2006).

The odds of a sentence being judged as having the most natural rhythm are defined as the probability of this event occurring divided by the probability of this event not occurring (Field 2005):

$$\text{odds}(\text{sentence has most natural rhythm}) = \frac{P(\text{sentence has most natural rhythm})}{1 - P(\text{sentence has most natural rhythm})}$$

These equations can be reformulated as follows (see Garson 2009) ('ln' means the natural logarithm of the odds, or 'log odds': Howell 2007).

$$\ln(\text{odds}(\text{sentence has most natural rhythm})) = b_0 + b_{dur}X_{dur} + b_{f0}X_{f0} + Z_i + \varepsilon_i$$

So the b-coefficients tell us the weighting of each variable in reference to the log odds of a sentence being judged as having the most natural rhythm. In Table 6-4, b_{dur} and b_{f0} are negative, i.e. durational and tonal deviance *decrease* the log odds of a stimulus being judged as most rhythmically-natural (positive values would indicate a log-odds increase). To see the relative contribution of each variable on the odds, which is easier to conceptualise than the log odds (cf. Garson 2009, Howell 2007), we exponentiate the b-coefficients, or raise e (the base of natural logarithms) to the power of b (Howell 2007), as Table 6-5 shows. We should not compare values here between languages (see footnote 2). Instead concentrate on the differential effects of durational and tonal deviance within languages: for a sentence to sound rhythmically-natural, non-deviant duration is far more important than non-deviant f0 for SGs (13.9% versus 1.5%), whereas non-deviant duration and non-deviant f0 are almost equally important for (S)Fr, though duration relatively less for SFr (20.6% versus 23.5%) and relatively more for Fr (29.2% versus 26.2%). This reinforces the conclusion that universally applied rhythm metrics do not capture what is perceptually significant for rhythm in different languages.

Variable (X)	Language	Standardised b-coefficient (sdX x b)	exp(b) (standardised)	This means that...
dur	SG	-0.150	0.861	if duration deviates from appropriate by 1sd , the odds of a stimulus being judged as most rhythmically-natural decrease by... 1-0.861=0.139 = 13.9%
	SFr	-0.231	0.794	1-0.794=0.206 = 20.6%
	Fr	-0.346	0.708	1-0.708=0.292 = 29.2%
f0	SG	-0.015	0.985	if f0 excursion deviates from appropriate by 1sd , the odds of a stimulus being judged as most rhythmically-natural decrease by... 1-0.985=0.015 = 1.5%
	SFr	-0.268	0.765	1-0.765=0.235 = 23.5%
	Fr	-0.304	0.738	1-0.738=0.262 = 26.2%

Table 6-5 – Explanation of standardised b-coefficients

Although exponentiated b-coefficients help to make sense of the regression output (by being relative to odds, not log odds), they have a drawback if used mathematically beyond this, since there is an asymmetry between a decrease in odds (varying only from 0 to 0.999...) and an increase in odds (varying from 1.001 to infinity) (Garson 2009). Therefore, the X-standardised non-exponentiated b-coefficients, after conversion to proportions of 1 in each language, were the weighting values used in the weighted PVIs (see Table 6-6).

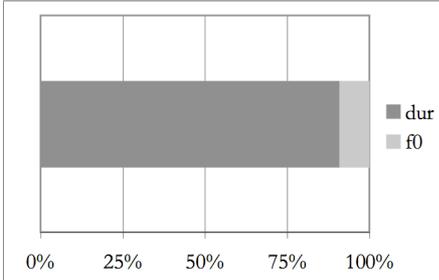
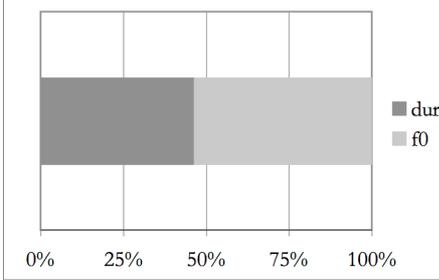
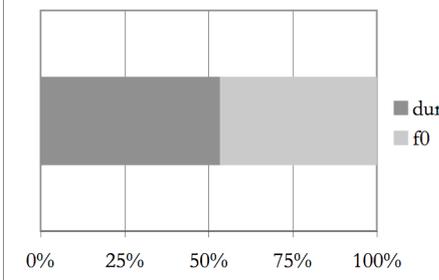
Language	$\frac{b_{dur}}{b_{f0}}$	$\frac{b_{dur}}{b_{f0}}$ (if $b_{dur}+b_{f0}=1$)	Relative weighting of $b_{dur} : b_{f0}$
SG	$\frac{-0.150}{-0.015}$	$\frac{0.908}{0.092}$	
SFr	$\frac{-0.231}{-0.268}$	$\frac{0.463}{0.537}$	
Fr	$\frac{-0.346}{-0.304}$	$\frac{0.533}{0.467}$	

Table 6-6 – Weighting values used in weighted PVIs (third column from left)

6.4.3.3 PVI calculations

Now to put these weighting values into PVIs, for which the formula is⁴:

$$\text{normalised PVI} = 100 \times \left[\sum_{k=2}^n \left| \frac{v_k - v_{k-1}}{(v_k + v_{k-1})/2} \right| / (n-1) \right]$$

n = number of intervals
(vowel/syllable)
 v = value of property p (duration/
f0 excursion) for k^{th} interval

⁴ Nolan and Asu (2009) suggested the notation v instead of d (for duration). d has been used in almost all publications featuring the PVI.

This means that the difference in an acoustic property between the members of successive pairs of intervals was calculated, then normalised by taking each difference as a proportion of the mean value within the pair, and averaged across the total number of interval-pairs in the speech analysed (Nolan and Asu 2009: 65). As in previous studies (e.g. Grabe and Low 2002, White and Mattys 2007a), some pairwise comparisons were over a pause. Figure 6-1 shows that the acoustic properties were:

- duration (in ms) for durational PVIs;
- f0 excursion (in st) for tonal PVIs;
- duration and f0 excursion in equal proportions (50:50) for combined PVIs;
- duration and f0 excursion weighted according to the language-specific weighting values in Table 6-6 for weighted PVIs. (Excursion, as opposed to any other f0 measurement, was necessary because the measurement had to correspond to that from which the f0 weighting values were derived, i.e. f0 excursion manipulations in experiment 3's stimuli.)

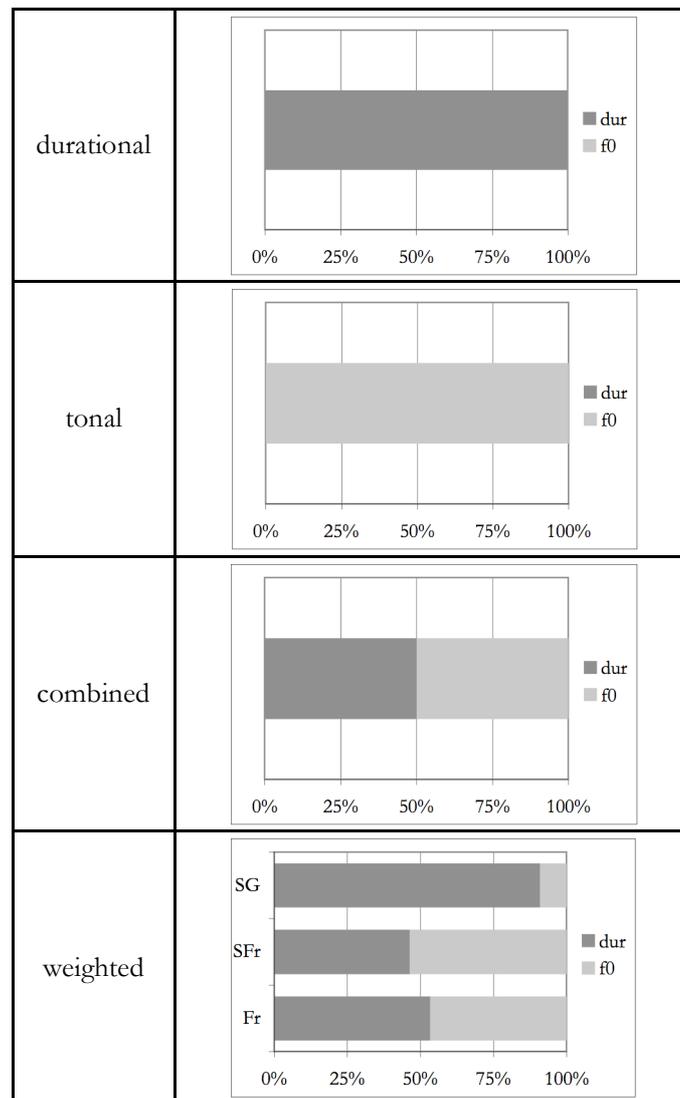


Figure 6-1 – Relative weighting of duration and f0 excursion for four PVI types

Table 6-7 is a step-by-step example of how each PVI type (for syllable variability) was calculated, using a six-syllable long utterance of a hypothetical language which shows high variability in syllable duration and f0 excursion. A summary of what the formulae represent appears at the top of each section.

Durational PVI				
1. Calculate the difference in duration between successive syllable-pairs. 2. Calculate the mean duration of successive syllable-pairs. 3. Divide each absolute pairwise difference (1.) by its respective pairwise mean (2.) to obtain each normalised pairwise difference. 4. Sum all the normalised pairwise differences (3.) from the utterance. 5. Divide the sum of normalised pairwise differences (4.) by the total number of syllables minus one, and multiply by 100.				
Syllable	duration (ms)	(1.) pairwise difference $v_k - v_{k-1}$	(2.) mean value of pair $(v_k + v_{k-1})/2$	(3.) normalised pairwise difference $\left \frac{v_k - v_{k-1}}{(v_k + v_{k-1})/2} \right $
1	50			
2	100	50	75	0.67
3	50	50	75	0.67
4	100	50	75	0.67
5	50	50	75	0.67
6	100	50	75	0.67
(4.) sum of normalised pairwise differences $\sum_{k=2}^n \left \frac{v_k - v_{k-1}}{(v_k + v_{k-1})/2} \right $				3.33
(5.) normalised PVI $100 \times \left[\sum_{k=2}^n \left \frac{v_k - v_{k-1}}{(v_k + v_{k-1})/2} \right / (n-1) \right]$				66.67
Tonal PVI				
6. Follow steps 1.-5. above, except use f0 excursion measurements.				
Syllable	f0 excursion (st)	(1.) pairwise difference $v_k - v_{k-1}$	(2.) mean value of pair $(v_k + v_{k-1})/2$	(3.) normalised pairwise difference $\left \frac{v_k - v_{k-1}}{(v_k + v_{k-1})/2} \right $
1	0.5			
2	5.0	4.5	2.75	1.64
3	0.5	4.5	2.75	1.64
4	5.0	4.5	2.75	1.64
5	0.5	4.5	2.75	1.64
6	5.0	4.5	2.75	1.64
(4.) sum of normalised pairwise differences $\sum_{k=2}^n \left \frac{v_k - v_{k-1}}{(v_k + v_{k-1})/2} \right $				8.18
(5.) normalised PVI $100 \times \left[\sum_{k=2}^n \left \frac{v_k - v_{k-1}}{(v_k + v_{k-1})/2} \right / (n-1) \right]$				163.64
Table continued overleaf				

Combined PVI (durational and tonal variability weighted 50:50)			
7. Calculate the normalised pairwise difference for duration and f0 excursion as in steps 1.-3. above. 8. Calculate, for each syllable-pair, the mean of the normalised pairwise differences for duration and f0 excursion to obtain the 'combined' $[(d+f)/2]$ pairwise difference. 9. Sum all the combined normalised pairwise differences (8.) from the utterance. 10. Divide the sum from step 9. by the total number of syllables minus one, and multiply by 100.			
Syllable	(3.)(7.) normalised pairwise difference for duration $\left \frac{d_k - d_{k-1}}{(d_k + d_{k-1})/2} \right $	(3.)(7.) normalised pairwise difference for f0 excursion $\left \frac{f_k - f_{k-1}}{(f_k + f_{k-1})/2} \right $	(8.) combined (c) normalised pairwise difference of duration (d) and f0 (f) excursion $c_k = \left[\left \frac{d_k - d_{k-1}}{(d_k + d_{k-1})/2} \right + \left \frac{f_k - f_{k-1}}{(f_k + f_{k-1})/2} \right \right] / 2$
1			
2	0.67	1.64	1.15
3	0.67	1.64	1.15
4	0.67	1.64	1.15
5	0.67	1.64	1.15
6	0.67	1.64	1.15
	(9.) sum of combined pairwise differences $\sum_{k=2}^n c_k$		5.76
	(10.) normalised PVI $100 \times \left[\sum_{k=2}^n [c_k] / (n-1) \right]$		115.15
Weighted PVI (durational and tonal variability weighted by language-specific b-coefficients)			
11. Calculate the normalised pairwise difference for duration and f0 excursion as in steps 1.-3. above. 12. Calculate, for each syllable-pair, the weighted (w) normalised pairwise difference of duration (d) and f0 excursion (f), by multiplying the d normalised pairwise difference by its b-coefficient (b_{dur}), and the f normalised pairwise difference by its b-coefficient (b_{f0}). The b-coefficients add up to 1. 13. Sum all the weighted normalised pairwise differences (12.) from the utterance. 14. Divide the sum from step 13. by the total number of syllables minus one, and multiply by 100.			
Syllable	(3.)(11.) normalised pairwise difference for duration $\left \frac{d_k - d_{k-1}}{(d_k + d_{k-1})/2} \right $	(3.)(11.) normalised pairwise difference for f0 excursion $\left \frac{f_k - f_{k-1}}{(f_k + f_{k-1})/2} \right $	(12.) weighted (w) normalised pairwise difference of duration (d) and f0 excursion (f): $b_{dur}=0.8, b_{f0}=0.2$ $w_k = \left[\left(0.8 \left \frac{d_k - d_{k-1}}{(d_k + d_{k-1})/2} \right \right) + \left(0.2 \left \frac{f_k - f_{k-1}}{(f_k + f_{k-1})/2} \right \right) \right]$
1			
2	0.67	1.64	0.86
3	0.67	1.64	0.86
4	0.67	1.64	0.86
5	0.67	1.64	0.86
6	0.67	1.64	0.86
	(13.) sum of weighted pairwise differences $\sum_{k=2}^n w_k$		4.30
	(14.) normalised PVI $100 \times \left[\sum_{k=2}^n [w_k] / (n-1) \right]$		86.06

Table 6-7 – How to calculate each PVI (hypothetical data)

We see that data from one (hypothetical) language results in four different PVI scores. Table 6-8 compares PVIs (values displayed in the lower part of the table) calculated for utterances in other hypothetical languages. Obviously no such perfectly regular languages exist. However, the point is that the difference between languages 1 and 3, and between 2 and 4 is not captured by purely durational PVIs, but *is* captured by the weighted PVIs, which (unlike combined PVIs) reflect the language-specific nature of perceived rhythm. (Each PVI was calculated as in Table 6-7; the weighting values are also fictitious.)

Language 1 Regular alternation between long and short syllables, and small and large f0 excursion.			Language 2 No variation in duration; regular alternation between small and large f0 excursion.			Language 3 No variation in f0 excursion; regular alternation between long and short syllables.			Language 4 No variation in duration or f0 excursion.		
syllable	duration (ms)	f0 excursion (st)	syllable	duration (ms)	f0 excursion (st)	syllable	duration (ms)	f0 excursion (st)	syllable	duration (ms)	f0 excursion (st)
1	50	0.5	1	75	0.5	1	50	0.5	1	75	0.5
2	100	5.0	2	75	5.0	2	100	0.5	2	75	0.5
3	50	0.5	3	75	0.5	3	50	0.5	3	75	0.5
4	100	5.0	4	75	5.0	4	100	0.5	4	75	0.5
5	50	0.5	5	75	0.5	5	50	0.5	5	75	0.5
6	100	5.0	6	75	5.0	6	100	0.5	6	75	0.5
Weighting values (b-coefficients expressed as proportions of 1)											
	0.8	0.2		0.2	0.8		0.9	0.1		0.5	0.5
PVIs											
	durational	66.67		durational	0.00		durational	66.67		durational	0.00
	tonal	163.64		tonal	163.64		tonal	0.00		tonal	0.00
	combined	115.15		combined	81.82		combined	33.33		combined	0.00
	weighted	86.06		weighted	130.91		weighted	60.00		weighted	0.00

Table 6-8 – Comparison of PVIs for hypothetical languages

6.5 Results

The following sections report the SG and (S)Fr PVI results, and other analyses of these subjects' speech data which are relevant to points that will emerge in the discussion (§6.6).

6.5.1 PVIs

Figure 6-2 plots the mean PVI scores, with vowel and syllable PVIs on separate graphs. The horizontal and vertical lines are the four axes (one for each PVI type) radiating out from a

centre point, and the faint grey regular quadrilaterals show steps of 20 along each dimension, up to 80. The Fr (red) quadrilateral fits inside the SFr (green) quadrilateral which fits inside the SG (blue) quadrilateral, for vowels and syllables, i.e. SG PVI > SFr PVI > Fr PVI (except vowel durational PVI where Fr is 0.39 higher than SFr). Each quadrilateral is a slightly different shape, i.e. the difference between PVI types varies cross-linguistically. No language has (like the grey lines joining the axes) an equilateral quadrilateral, particularly as tonal PVIs are higher than the others. The shaded areas are smaller for syllables than vowels, i.e. vowel PVI > syllable PVI.

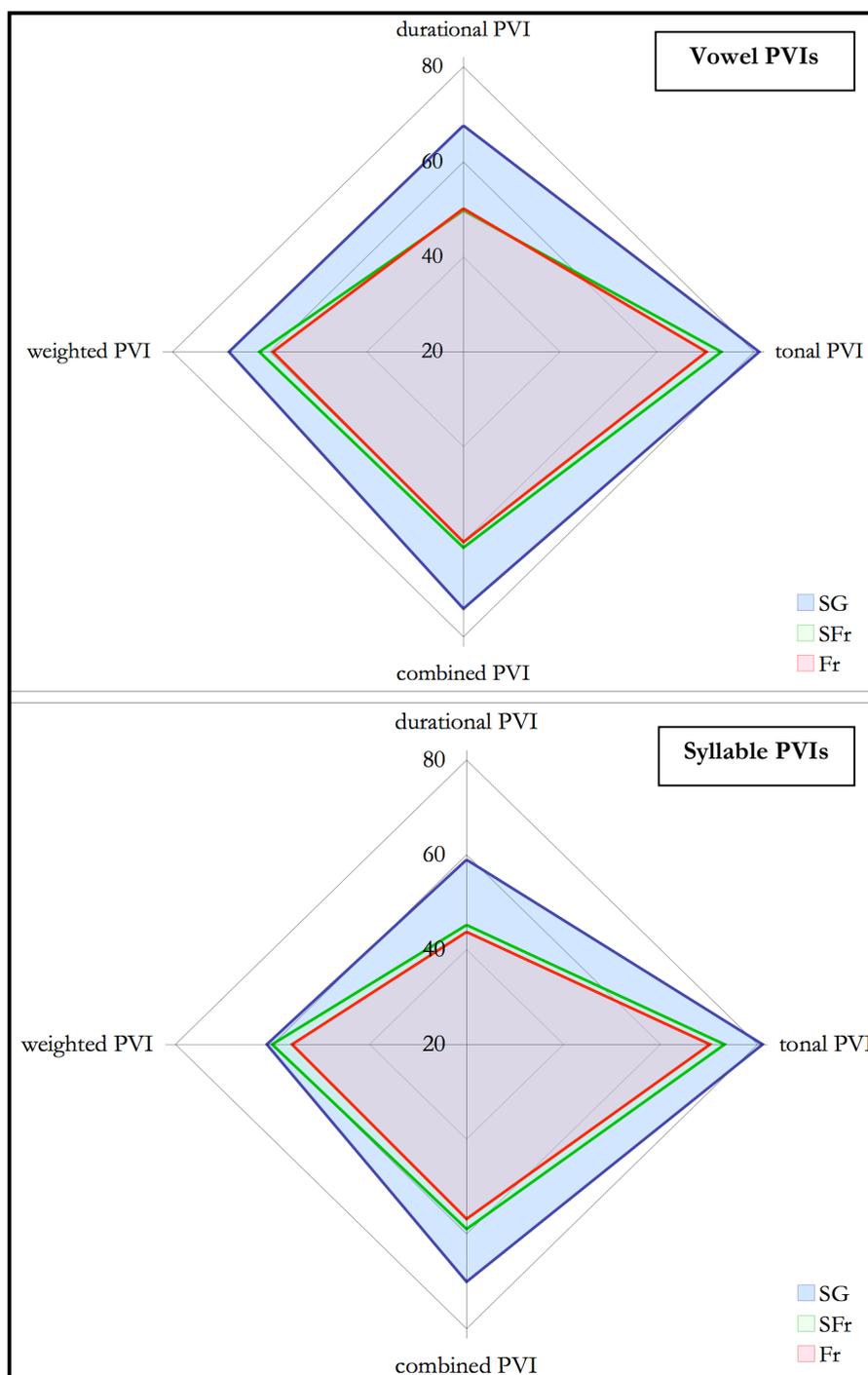


Figure 6-2 – Mean PVI scores (10 subjects per language); tonal PVI data are identical in both graphs (see §6.4.3.1)

Before these PVI scores were input to an ANOVA in *SPSS*, they were explored for normality and homogeneity of variance. No data-series were significantly non-normal according to Shapiro-Wilk tests ($p > 0.05$) and visual inspection of histograms. According to Levene tests, the data-series for vowel and syllable tonal PVIs, but no others, significantly violated the homogeneity of variance assumption ($p < 0.01$), but as sample sizes were equal and ANOVA is relatively robust against assumption violations (Howell 2007), the data were not transformed. The ANOVA was three-way and mixed-measures, with the factors *interval* (vowel, syllable), *PVI-type* (durational, tonal, combined, weighted) (both repeated-measures) and *language* (SG, SFr, Fr). All main effects and interactions were significant (Table 6-9).

Source	df	Mean square	F	p
interval	1	724.121	100.056	<0.0001***
interval × language	2	38.852	5.368	0.011*
Error	27	7.237		
PVI-type	1.084	13876.688	265.931	<0.0001***
PVI-type × language	2.169	411.214	7.880	0.001**
Error	29.280	52.182		
interval × PVI-type	1.226	271.540	91.795	<0.0001***
interval × PVI-type × language	2.452	18.089	6.115	0.003**
Error	33.103	2.958		
language	2	409.264	30.594	<0.0001***
Error	27	13.377		
<ul style="list-style-type: none"> • A Greenhouse-Geisser correction was applied since sphericity could not be assumed (Mauchly's test, $p < 0.001$) • significance: *** $p < 0.0001$; ** $p < 0.01$; * $p < 0.05$ 				

Table 6-9 – ANOVA output: *interval* × *PVI-type* × *language*

The main effect of *interval* confirms that vowel PVIs were significantly higher than syllable PVIs. Post-hoc tests (Tukey HSD) for *language* showed a significant difference between SG and SFr, and SG and Fr ($p < 0.0001$), but not between SFr and Fr ($p > 0.05$). Planned comparisons between durational PVIs and each of the other PVIs explored the main effect of, and interactions with, *PVI-type* (Table 6-10).

Source		df	Mean square	F	p
PVI-type	durational vs. tonal	1	14773.537	311.459	<0.0001**
	durational vs. combined	1	3600.362	304.729	<0.0001**
	durational vs. weighted	1	2208.634	613.597	<0.0001**
PVI-type × language	durational vs. tonal	2	165.192	3.483	0.045*
	durational vs. combined	2	38.979	3.299	0.052 ^m
	durational vs. weighted	2	400.683	111.317	<0.0001**
Error	durational vs. tonal	27	47.433		
	durational vs. combined	27	11.815		
	durational vs. weighted	27	3.599		
interval × PVI-type	durational vs. tonal	1	1286.113	100.194	<0.0001**
	durational vs. combined	1	350.996	101.642	<0.0001**
	durational vs. weighted	1	162.328	51.518	<0.0001**
interval × PVI-type × language	durational vs. tonal	2	42.401	3.303	0.052 ^m
	durational vs. combined	2	14.172	4.104	0.028*
	durational vs. weighted	2	6.197	1.967	0.159
Error	durational vs. tonal	27	12.836		
	durational vs. combined	27	3.453		
	durational vs. weighted	27	3.151		
significance: ** $p < 0.0001$; * $p < 0.05$; ^m marginally significant, $p < 0.1$					

Table 6-10 – Planned comparisons within ANOVA (*interval* × *PVI-type* × *language*)

Without accounting for *interval* or *language*, durational PVIs were significantly lower than all other PVI types ($p < 0.0001$). If we consider the *interval* × *PVI-type* interaction, the difference between durational PVIs and each of the other PVIs was significantly greater for syllable than vowel variability ($p < 0.0001$). If we consider the *PVI-type* × *language* interaction, the difference between durational and tonal PVIs was significantly greater for (S)Fr than SG ($p = 0.045$), the difference between durational and combined PVIs was almost significantly greater for (S)Fr than SG ($p = 0.052$), and the difference between durational and weighted PVIs was significantly greater for (S)Fr than SG ($p < 0.0001$). On Figure 6-2, these significant differences appear as a larger blue area beyond the green and red lines on the durational axis (from the centre northwards)

compared to the smaller blue area beyond the green and red lines eastwards, westwards and to some extent southwards. If we take into account *language* and *interval* when comparing PVI types (three-way interaction), only combined PVIs were significantly different from durational PVIs, though the tonal versus durational PVI difference just failed to reach significance at $p < 0.05$. This means that the difference between durational and weighted PVIs can be explained as the effects of *interval* (vowel > syllable PVIs) and *language* (SG > (S)Fr PVIs).

As the predictions detailed, it is unsurprising that SG displays more pairwise durational and tonal variability than (S)Fr. More interesting is the result that SFr is generally more durationally and tonally variable than Fr, particularly for syllables (i.e. in Figure 6-2 the Fr quadrilateral fits neatly into the SFr quadrilateral). Another three-way ANOVA with the factors *interval*, *PVI-type* and *language* was run to compare just SFr and Fr. Main effects occurred for *interval* [$F(1,18)=42.671$, $p < 0.0001$] and *PVI-type* [$F(1.008,18.148)=607.791$, $p < 0.0001$] but not *language* [$F(1,18)=3.056$, $p > 0.05$], and the *interval* \times *PVI-type* interaction (but none other) was significant [$F(1.017,18.306)=41.605$, $p < 0.0001$]. Planned comparisons between durational and each of the other PVIs explored the main effect of, and interactions with, *PVI-type*. From the bolded cells in Table 6-11, we see that the difference between durational and weighted PVIs was significantly greater for SFr than Fr when averaged across *interval* (*PVI-type* \times *language*, $p=0.009$), i.e. in Figure 6-2 the green area beyond the red line is larger westwards than northwards. When the variation explained by *interval* (vowel > syllable PVIs) is included, this significant difference between SFr and Fr disappears (*interval* \times *PVI-type* \times *language*, $p=0.595$). This means that subtle differences between SFr and Fr rhythm may be reflected in weighted as opposed to durational PVIs, but also depending on the speech unit measured (i.e. syllable/vowel).

Source		df	Mean square	F	p
PVI-type	durational vs. tonal	1	11929.557	610.312	<0.0001**
	durational vs. combined	1	2902.260	574.614	<0.0001**
	durational vs. weighted	1	2946.500	604.108	<0.0001**
PVI-type × language	durational vs. tonal	1	31.567	1.615	0.220
	durational vs. combined	1	6.503	1.288	0.271
	durational vs. weighted	1	42.029	8.617	0.009*
Error	durational vs. tonal	18	19.547		
	durational vs. combined	18	5.051		
	durational vs. weighted	18	4.877		
interval × PVI-type	durational vs. tonal	1	600.883	41.980	<0.0001**
	durational vs. combined	1	154.402	41.031	<0.0001**
	durational vs. weighted	1	152.229	38.570	<0.0001**
interval × PVI-type × language	durational vs. tonal	1	16.582	1.158	0.296
	durational vs. combined	1	3.615	0.961	0.340
	durational vs. weighted	1	1.159	0.294	0.595
Error	durational vs. tonal	18	14.314		
	durational vs. combined	18	3.763		
	durational vs. weighted	18	3.947		
significance: ** $p < 0.0001$; * $p < 0.01$					

Table 6-11 – ANOVA output: *interval × PVI-type × language* (SFr vs. Fr)

6.5.2 Other analyses

The duration and f_0 measurement data from which the PVIs were calculated were used for further analyses. First, the extent to which the magnitude of durational difference between individual vowels correlated with that of tonal difference was found, to explore how precisely increased duration and large f_0 movements co-occur in SG and (S)Fr, as the predictions (§6.3) suggested. Second, subjects' speech rate was calculated, to explore whether SFr and Fr speakers differed, since anecdotal reports describe SFr as slower than Fr. These analyses are now presented in turn, and will be relevant in later discussion (§6.6).

In one correlation analysis per subject, two data-series were compared against one another: for each vowel-pair, the normalised between-vowel difference in duration versus the normalised between-vowel difference in f0 excursion. Each data-point was equivalent to:

$$\frac{|v_k - v_{k-1}|}{(v_k + v_{k-1})/2} \quad v = \text{duration or f0 excursion of } k^{\text{th}} \text{ vowel}$$

Table 6-12 shows that for half the subjects there was no significant correlation of these durational differences and tonal differences. For the half that showed a significant correlation, the highest coefficients were no greater than 0.4, still far from a perfect linear relationship ($r=1$). Although increased duration and larger f0 excursion may generally co-occur, this lack of, or at best weak, correlation suggests that these two dimensions vary quite independently of each other.

	Subject	Coefficient (r)	Total number of vowels – 1 (n)	Significance (p)
SG	AS	0.2785	130	<0.01
	BS	0.1970	129	<0.05
	CEF	0.0281	130	non-sig.
	MB	0.1577	131	non-sig.
	SMF	0.3931	129	<0.001
	AH	0.0445	127	non-sig.
	CEM	0.0280	130	non-sig.
	EZ	0.0769	129	non-sig.
	RM	0.2131	129	<0.05
	SMM	0.1892	134	<0.05
SFr	CM	0.1502	162	non-sig.
	EC	0.1396	164	non-sig.
	JP	-0.0166	157	non-sig.
	JV	0.0807	163	non-sig.
	SW	-0.0250	164	non-sig.
	BP	0.3952	163	<0.001
	CR	0.2414	160	<0.01
	CW	0.2207	167	<0.01
	LC	0.2668	160	<0.001
	PG	-0.0283	159	non-sig.
Fr	AH	0.1787	158	<0.05
	BD	-0.1662	160	<0.05
	FP	0.0267	155	non-sig.
	LD	-0.0334	163	non-sig.
	NM	0.0223	163	non-sig.
	FB	0.1755	145	<0.05
	LP	0.1911	158	<0.05
	OA	0.1791	155	<0.05
	VD	0.2229	145	<0.01
	VG	0.0649	150	non-sig.

Table 6-12 – Correlation output (per subject)

Several Fr subjects commented that they thought their own natural speech rate was fast. Given this, and previous anecdotes that SFr is slow, speech rate was calculated (from the data shown in Table 6-13) and compared across language groups. A one-way ANOVA was run for each rate-related property; planned comparisons compared Fr with SFr and Fr with SG.

Measure per recording		SG	SFr	Fr	One-way ANOVA results
Total duration (secs)	\bar{x}	32.8	33.8	31.6	$F(2,27)=1.406, p>0.05$
	sd	3.3	2.4	2.9	
Total duration – pause durations (secs)	\bar{x}	25.7	27.3	24.6	$F(2,27)=4.194, p=0.026^*$ Fr vs. SFr: $p=0.008^{**}$ Fr vs. SG: $p>0.05$
	sd	2.2	1.8	2.3	
Number of phonologically underlying syllables		134	169	169	
Number of syllables elided relative to underlying number ^a	\bar{x}	3.3	6.1	11.0	$F(2,27)=8.562, p=0.001^{**}$ Fr vs. SFr: $p>0.05$ Fr vs. SG: $p<0.0001^{***}$
	sd	1.6	2.9	6.5	
Rate (syllables/sec) (excluding pause time)	\bar{x}	5.1	6.0	6.4	$F(2,27)=24.443, p<0.0001^{***}$ Fr vs. SFr: $p=0.04^*$ Fr vs. SG: $p<0.0001^{***}$
	sd	0.4	0.4	0.5	
<ul style="list-style-type: none"> • significance: $^{***} p<0.0001$; $^{**} p<0.01$; $^* p<0.05$ • ^a In these languages elision concerned vowels, but if a vowel was elided, any consonants were resyllabified with the surrounding syllables so vowel elision equated to syllable elision. 					

Table 6-13 – Speech-rate-related properties measured from recordings

From the standard deviations (sd), we see that inter-speaker variation occurred within languages/varieties, particularly for vowel elision, which was much more variable between speakers within Fr (sd=6.5) than within SFr (sd=2.9) and SG (sd=1.6). Between language groups, total duration minus pause time was significantly higher for SFr than Fr, but similar for SG and Fr. Vowel elision was much more frequent in Fr than in SFr and SG, though only the between-language (not between-variety) contrast was significant. More frequent vowel elision meant that Frs produced fewer syllables, and had a significantly faster mean rate than SGs and SFr.

6.6 Discussion

6.6.1 (Vowel or syllable) durational PVIs

For (normalised) durational PVIs, it was predicted that SG would be relatively high and (S)Fr relatively low within the range of previously reported values. The results provide evidence for this, as vowel durational PVIs were similar values to those in previous studies (e.g. Galloway 2007, Grabe and Low 2002, Lee and Todd 2004, Schmid 2001, White and Mattys 2007a), with

SG significantly higher than (S)Fr, and no significant difference between SFr and Fr. Syllable durational PVIs, though much lower than vowel PVIs, were also significantly different between languages but not language-varieties.

A possible explanation for different vowel and syllable variability is the differential effect on vowels and consonants of reduction in connected speech. In SG, all non-prominent (non-loanword) syllables contain [ə] or [i] which are reduced and central(ised) (Fleischer and Schmid 2006, Reese 2007). In (S)Fr, optional schwas are produced to maintain clusters of maximally two consonants, though larger clusters may occur if optional schwas are suppressed, often in faster speech (see §6.4.3.1). In this experiment, many speakers (both languages) sometimes produced very short (a trace two or three glottal periods long in the acoustic record) [ə] or [i] in non-prominent syllables, for example: (S)Fr [sə.di.spy.tɛ] ('were arguing'), [ki.sa.vã.sɛ] ('who was approaching'), i.e. not quite full suppression of schwa and [i] in function words; SG ['ti.kʰə] ('thick'), ['bri.ŋi] ('to bring'). The consonants in these syllables, e.g. [k] [s] [kʰ] [ŋ], were not markedly reduced compared to the vowels. Logically then, vowel variability is greater than syllable variability, because syllable variability also includes consonants, which are less durationally variable than vowels.

This experiment measured vowels and syllables to compare, in the same data-set, these acoustic/phonetic and phonological approaches to measuring durational and tonal variability. If we measure produced rhythm (in SG and (S)Fr) acoustically, as though listening through infant ears, successive intervals (vowels) show relatively high acoustic variability (also depending on language). Therefore, rhythm unconnected to phonological knowledge of a language (babies hearing speech) may have a relatively clear alternation of perceptually 'weak' and 'strong' elements. If we measure rhythm phonologically, as though listening under the influence of knowledge acquired by adulthood, successive intervals (syllables) show significantly lower acoustic variability than vowels. Therefore, rhythm connected to phonological knowledge of a language (adults hearing speech) may have a perceptually less clear 'weak'–'strong' pattern. This reduced clarity of rhythm results from the fact that meaningful speech is an acoustically and linguistically complex signal with several interacting factors determining its properties at any point. Different approaches to measuring rhythm, e.g. Grabe and Low (2002) (phonetic) and Nolan and Asu (2009) (phonological), might have measured phenomena that are to some extent perceptually distinct. Vowel (and consonant) variability is inextricably linked to syllable variability, but syllable variability is arguably a more appropriate measure if we are concerned with adults' perception of rhythm in language, because they have knowledge of their native-language phonology including syllables (and larger prosodic groups) (cf. Nolan and Asu 2009).

6.6.2 Tonal PVIs

No previous studies have calculated tonal PVIs. It was predicted that, for each language, (normalised) tonal PVIs would be similar to durational PVIs, because lengthening and f_0 movement co-occur in SG and (S)Fr. This prediction was not supported, since tonal PVIs were significantly higher than durational PVIs. (Since the PVIs were all normalised, the difference in measurement units between ms and st is irrelevant.) F_0 excursion was chosen as the unit measure of pitch because, for calculating weighted PVIs, the f_0 weighting values had to be multiplied by an excursion value (see §6.4.3.2). A set of tonal PVIs was also calculated with f_0 velocity as input, which accounted for the possibility that longer vowels have larger excursion (as they have ‘more room’ for pitch movement) than shorter vowels. These PVIs were also much higher than durational PVIs, so it was decided that f_0 excursion should remain the measurement for tonal, combined and weighted PVIs. A possible explanation for higher tonal than durational PVIs is that tonal movements tended to be either large (i.e. pitch-accents) or microfluctuations less than 1st, whereas shorter vowels, though (of course) shorter than longer ones, still had a considerable ms value. For instance, adjacent vowels could have a difference of 150–75ms and 5–1st, a ratio of 2:1 for length but 5:1 for pitch. Normalised PVIs account for differences within one measurement unit, e.g. speaking rate (duration) or pitch range (f_0 excursion), but do not cover up this potentially important finding that durational variability was of lower magnitude than tonal variability.

In §6.3.2, some factors were suggested that might result in tonal PVIs lower and higher than durational PVIs in SG and (S)Fr respectively. Although tonal PVIs were higher overall than durational PVIs, the difference between tonal and durational PVIs was significantly lower in SG than in (S)Fr; therefore, if we acknowledge that tonal variability was always higher than durational variability (for the reason just explained), there is some evidence for these factors. In SG, adjacent syllables may ‘share’ a late-starting pitch-accent (see chapter 2), so tonal variability was not much higher than the high durational variability. In (S)Fr, speakers may realise rhythmic-group-initial prominence with substantial f_0 movement but not increased duration as in group-final prominence (see chapter 2), so tonal variability was much higher than the low durational variability. Tonal PVIs were marginally higher in SFr than Fr, perhaps indicating that SFr group-initial and group-final f_0 movements occurred closer together than in Fr, hence more evenly spaced large excursions.

In fact, the magnitude of between-vowel durational differences did not correlate perfectly with that of tonal differences, for any language (§6.5.2). Prominent syllables were longer than surrounding non-prominent ones, but the precise durational difference between them depended on many factors, e.g. each syllable’s structure and the inherent length of its segments. Prominent syllables had larger f_0 excursion than surrounding non-prominent ones, but the precise tonal difference between them depended on many factors, e.g. the number and length of

syllables between phonologically specified tones, and inter-speaker variation in the phonetic timing of f_0 movements. Speakers read the same text, but could choose to convey the meaning in various ways by emphasising certain parts over others, which affected the durational and tonal characteristics of individual syllables. This imperfect overlap of durational and tonal variability demonstrates some of the acoustic complexity of the speech signal. These languages are more acoustically variable than their durational PVIs capture, which is an important point, given that tonal variability is involved in rhythm perception (chapter 5).

6.6.3 Weighted PVIs

From experiment 3's findings (chapter 5), it was predicted that in SG duration would have greater weighting than f_0 , whereas in (S)Fr the cues would have more equal weighting. This was confirmed in the calculation of weighting values, which also showed that f_0 was marginally more important than duration for SFr, and the opposite for Fr (note that SFr listeners heard Fr in the perceptual experiment). As suggested in chapter 5, it seems that SG speakers are most sensitive to the obvious durational variability of SG, whereas (S)Fr speakers may be sensitive to a more even durational and tonal pattern across syllables, when they perceive rhythm in their native language. SFr's slightly greater sensitivity to f_0 than Frs may result from the difference in precise timing and placement of pitch movements in SFr and Fr (see chapter 5 and Miller 2007).

Consequently, it was predicted that SG durational and weighted PVIs would be similar, whereas (S)Fr weighted PVIs would lie between durational and tonal ones. The results support these predictions, since (S)Fr weighted PVIs (averaged over vowel and syllable PVIs) were significantly higher than durational PVIs and lower than tonal PVIs, whereas SG durational and weighted PVIs were almost identical. However, when the fact that vowel PVIs were higher than syllable PVIs was accounted for, the difference between durational and weighted PVIs (when compared across languages) was not as significant as the difference between tonal or combined and durational PVIs. Two important points follow from these findings.

First, the weighted PVIs were a compromise between separate measures of variability for two acoustic cues which turned out to be very different. Viewed in isolation, durational variability was much lower than tonal variability, as explained above, though both are relevant to perceived rhythm. Weighted PVIs combined these two sets of different but related information, thus capturing two dimensions of the multi-dimensionality of rhythm in a complex speech signal. In fact, unlike combined (non-weighted) PVIs, weighted PVIs captured another 'dimension': the language-specific perceptual relevance of each cue. Thus the difference between combined PVIs and weighted PVIs was lower in (S)Fr, with fairly equally weighted cues, than in SG, with length weighted much higher than pitch for perceived rhythmicity.

Second, the magnitude of cross-linguistic difference shown by the PVIs was as follows: weighted PVIs < tonal PVIs < combined PVIs < durational PVIs (in Figure 6-2, the smaller the

cross-linguistic difference, the closer the blue, green and red lines). Therefore, when we quantify rhythm taking into account its multi-dimensionality and the language-/variety-specific relevance of its cues, we find that languages usually classified as rhythmically distinct are more similar than universal durational metrics demonstrate. An illustration of what this might mean in terms of cross-linguistic differences in rhythm is as follows. To a group of listeners with a particular native language (say, English) the rhythm of some other languages (say, Spanish) seems different. However, if English listeners could hear Spanish ‘through the ears’ of Spanish native speakers, i.e. have native-like knowledge of Spanish phonology including the relative importance of various rhythm cues, these English listeners might think that the phenomenon of rhythm shows much similarity between these languages. Durational metrics apparently quantify cross-linguistic differences in rhythm, but these metrics might inappropriately capture cross-linguistic differences that are less significant in rhythm *perceived* by native speakers of a language than the similarities captured by a multi-dimensional language-specific metric.

6.6.4 Inter-speaker variation

Rhythm production studies have rarely discussed inter-subject variation. Wallin (1901: 8) recognised the need for a mixed subject-pool, and recorded subjects ‘with different languages, [...] of different stages and walks of life ([including] [...] school pupils, [...] students, professors, poets, orators, musicians etc.)’. Grabe and Low (2002) measured data from only one speaker of each language that they investigated. However, Grabe (2002) admitted that ‘comparable data from several speakers of several dialects of each of a number of languages’ should be collected before legitimate cross-linguistic comparisons concerning rhythm typology could be made. This was in light of some Spanish PVI data, in which some inter-speaker differences were as large as the cross-linguistic differences found by Grabe and Low (2002). Similarly, Barry et al. (2003) and Schmid (2004) found that various Italian dialects had some significantly different rhythm-metric scores. Kohler (2009a), quoting Classe (1939) and Cicero, made the point that some speakers, good orators, are just better than others at producing rhythmical utterances.

The present experiment’s results show inter-speaker variation, within and between languages. Within-language variation is illustrated by the standard deviations (sd) of PVIs (appendix 8.4.2). For example, syllable weighted PVIs had a mean of 61.14, 60.02 and 55.93, and sd of 3.88, 2.33 and 3.12 in SG, SFr and Fr respectively. This means that about six to seven (of ten) speakers per language fell within the range 57.26–65.01, 57.69–62.35 and 52.82–59.05 (SG, SFr and Fr respectively); others were just outside these ranges, i.e. between the lowest and highest speakers per language there was a range of around 10 PVI points. The sd of some speech-rate-related measures (§6.5.2) was also relatively high, e.g. number of elided vowels (from the phonologically underlying number) for SFr and Fr. Since speakers were instructed to speak however they felt comfortable, it is unsurprising that rhythm (measured by PVIs) and speech rate

emerged differently in various individuals. Some speakers' natural speech style was probably faster/slower and more/less rhythmical than others' (though how we would define 'rhythmical' in terms of PVI is not clear). Normalised PVI controlled for variable speech rate.

Contrary to Simon's (2003) claim that non-Parisian Fr speakers tend to produce standardised 'reading prosody' (see §6.4.2), this experiment found variation within and between SFr and Fr, even though all speakers were reading. SFr PVI was generally (non-significantly) higher than Fr PVI, except vowel durational PVI (SFr=49.74, Fr=50.13). The magnitude of between-variety difference shown by the PVI was as follows: weighted PVI > tonal PVI > combined PVI > durational PVI (in Figure 6-2, the greater the between-variety difference, the larger the green area). (Note the complete reverse order for cross-linguistic difference shown by the PVI: §6.6.3.) Crucially, a between-variety difference was captured less clearly with durational PVI than when rhythm was measured multi-dimensionally and variety-specifically with weighted PVI. Nevertheless, the between-variety difference in weighted PVI depended on whether syllables or vowels were measured, and its magnitude was less than cross-linguistic differences in durational PVI. Speech rate was calculated to investigate anecdotal reports (previously and from these subjects) that Fr is spoken faster than SFr. Miller (2007), also from recordings of 'La bise et le soleil', found that SFr rhythmic groups and IPs were longer than in Fr, but when pause time was excluded, there was no significant between-variety difference in speech rate. Conversely in the present experiment, mean speech rate (excluding pause time) was significantly higher in Fr than SFr, supporting the anecdotal claims. On average, SFr had fewer elided vowels (from the underlying number) and a significantly slower rate than Fr, which accords with the fact that in (S)Fr, optional schwas are less likely to be produced in fast than in careful speech (see §6.4.3.1). (Miller's (2007) SFr speakers' longer phrases might have had fewer elided vowels than her Fr speakers'.) Between-variety differences in reduction did not translate to significantly different rhythms when quantified with durational PVI. When tonal characteristics and the cues' perceptual relevance were accounted for with weighted PVI, more of the subtle prosodic variation between SFr and Fr was captured.

As expected, all PVI differed significantly between SG and (S)Fr, reflecting their markedly different rhythmic structures. The speech-rate-related measures also demonstrated these languages' different linguistic structures. The texts conveyed the same meaning with fewer syllables in SG than Fr, perhaps because many SG function morphemes are syllabically indistinct from the noun or verb associated with them. For example, in 'd Sune' [tsu.nə] ('the sun') and 'gschritte' [kʃtri.tə] ('argued'), the definite article 'de' and the past participle morpheme 'ge-' have been reduced historically and are now grammaticalised as 'd-' and 'g-'. We might expect that having complex syllables allows SG to build words with fewer syllables, which might also contribute to that text's lower syllable count. The SG recordings had fewer syllables than the (S)Fr recordings, but a total duration (excluding pause time) similar to the Fr recordings, so SG

speakers had a significantly slower rate (syllables/sec) than Fr speakers. This is unlikely to result from SGs being less familiar with reading dialect than (S)Fr speakers reading Fr, since fluent recordings without hesitation were selected, and these sounded as natural as the (S)Fr recordings. This slower rate reflects the fact that many SG syllables are complex (see examples above) so are generally longer than (S)Fr syllables, which relates back to different impressions of rhythm.

6.7 Conclusion

The previous experiments (chapters 3-5) found that f_0 and duration are interdependent in the perception of isolated syllables, rhythmic groups and sentence rhythmicity, and that native language affects the relative weighting of these cues. Two major implications of these findings for duration-based rhythm research are that it should also investigate f_0 (plus potentially other acoustic cues), and consider that rhythm *perception* differs cross-linguistically, to avoid a ‘one-size-fits-all’ method when comparing cross-linguistic rhythmic differences (in production). The point of this experiment was to suggest a method for making a quantification of produced rhythm perceptually informed, by combining f_0 and duration and weighting these according to their language-specific significance in perceived rhythmicity. An evaluation of this suggestion is now needed. According to Nolan and Asu (2009), previous assessments of rhythm metrics have judged performance by how well they correlate with impressionistic rhythm ‘types’. Since the present experiment did not assume that a rhythm typology necessarily exists, it did not aim to better categorise these languages into ‘types’ by generating new numbers. Instead, we need to ask whether weighted PVIs are better than traditional durational ones in making it possible to state what the numbers mean in terms of perception (cf. Barry et al. 2009).

In speech, perceived rhythm is induced by a prominence pattern, which depends on how each syllable’s unique combination of duration and f_0 (and potentially other cues) differs from that of its neighbours, and on the relative perceptual significance of each cue in the language concerned. This acoustic complexity is not fully represented by durational (or indeed tonal) metrics. Weighted PVIs represent a language-specific compromise between durational and tonal variability, so they more adequately reflect the complex interaction of linguistic factors and acoustic properties that differs cross-linguistically and results in perceived rhythm in language. In sum, perceived rhythm is now represented in numbers: two acoustic cues and cross-linguistic variation in perception are involved. The weighted PVIs for SG and (S)Fr provide evidence that if we account for the acoustic multi-dimensionality and language-specificity of perceived rhythm when quantifying produced rhythm, the phenomenon might not be as cross-linguistically divergent as we think. Furthermore, the above discussion deliberately highlighted the within-language and between-variety variation observed in rhythm production that studies with a typological perspective have tended to ignore. This variation demonstrates the need to investigate

several subjects' rhythm production and perception, comparing within and between languages/varieties.

The weighting values for weighted PVIs were derived from adults' perception, who are aware of their native-language phonology. Therefore, weighted PVIs with syllables (i.e. phonologically defined intervals) mean more in terms of representing linguistic rhythm perception than weighted PVIs with vowels (i.e. acoustically defined intervals). A potential issue is that for syllable PVIs, f_0 excursion was only measured across the vowel, but for good theoretical and practical reasons: perceptually relevant f_0 movements occur in steady-state vowels (House 1990), and f_0 is lost during many consonants, neither of which affects duration. This is an inevitable consequence of the different acoustic nature of duration and pitch. Although the syllable appears to be the domain over which large f_0 movements (i.e. pitch-accents) occur in these languages, tonal variability could be measured between longer prosodic (e.g. rhythmic) groups, since these might also capture some of the variability relevant to perceived rhythmic structure.

Weighted PVIs have advanced from traditional durational metrics which told nothing of perceived rhythm, but the progress is just one small step. The positive evaluation above is based on data from only two languages, including two varieties of one. Wider research with many more prosodically diverse languages is needed to fully assess the usefulness of weighted PVIs. Other acoustic cues like amplitude and spectral properties are also likely to interact to some extent with duration and f_0 in rhythm perception (see chapter 1). Thus an even better representation of perceived rhythm would be PVIs that include language-specific weighting of more cues. The statistical method suggested for linking perceptual findings to production data, by developing an already popular method for quantifying rhythm, was a first attempt and has room for improvement. Currently a perceptual experiment like that reported in chapter 5 needs to be run to obtain language-specific cue weightings. Since a cross-linguistically extensive investigation is proposed, this would be a time-consuming task, though it could be shared by several researchers who could each adopt the same method to make results comparable. Large-scale cross-linguistic rhythm research linking perception and production should not be avoided just because it is challenging. In fact, this is precisely what is needed in light of the experiments within this thesis.

Conclusions

7.1 Contributions to speech-rhythm research

This thesis has contributed to research on speech rhythm by investigating the subject from three perspectives which current experimental phonetic research generally ignores. The aims of the thesis, given in chapter 1, reflected these perspectives:

- **A focus on perception.** The aim was to investigate perceived rhythm, to link the findings to the PVI, and make this metric perceptually informed.
- **The inclusion of f0.** The aim was to observe not just duration, but how f0 and duration are interdependent in rhythm perception and production.
- **A cross-linguistic study.** The aim was to provide evidence for or against native-language/-language-variety influence on perceived rhythm.

Table 7-1, which summarises the research questions and answers for each of the four experiments, shows that these aims have been achieved.

Experiment (chapter)	Research question	Evidence that answer is affirmative	Native language effect	Next step
1 (3)	Does a dynamic f0 affect the perceived duration of non-speech sounds and isolated monosyllables? If so, does this depend on listeners' native language?	A perceived lengthening effect of dynamic f0 was observed.	The effect was stronger for linguistic than non-linguistic stimuli, but did not depend on native language.	This finding in a psycho-acoustic/-phonetic task laid the basis for further investigation in a prosodic domain longer than syllable pairs.
2 (4)	Are dynamic f0 and increased duration interdependent perceptual cues to rhythmic groups? If so, does this depend on listeners' native language?	Two cues were more effective than one when heard simultaneously, but less effective than one when heard in conflicting positions around the rhythmic-group boundary.	Native language affected whether increased duration or dynamic f0 was the more effective cue.	The main finding extended that of experiment 1 to a longer prosodic domain. Thus an experiment addressing the interdependence of f0 and duration in rhythm perception was then worthwhile.
3 (5)	Are f0 and duration interdependent perceptual cues when listeners have to judge the rhythmicality of sentences? If so, does this depend on listeners' native language?	How (non-)deviant the duration <i>and</i> f0 excursion of prominent syllables were contributed to whether sentences were perceived as rhythmically-natural.	The relative weighting of a non-deviant duration and non-deviant f0 excursion depended on native language.	This was confirmation that both cues should feature in rhythm research. These perceptual data were then used in a production experiment.
4 (6)	Can we make the PVI perceptually informed by applying the perceptual data from experiment 3 to production data?	Weighted PVIs, quantifications of produced rhythm, captured the acoustic multi-dimensionality and language-specificity of perceived rhythm.		This seemed successful, but refinement of the method is welcome, and investigation of more languages is essential.

Table 7-1 – Summary of research questions and answers

The research questions for each experiment were more specific than the general research questions (outlined in chapter 1), which were as follows.

- Does f_0 play a significant role in speech-rhythm perception, and how does its significance compare with that of duration? Are tonal and durational cues interdependent?
- Are native speakers of different languages (and language-varieties) sensitive to durational and tonal rhythm cues to different extents? If so, is rhythm perception more language-(/variety-)specific than universal?

From the experiments' findings, we can now answer these general questions. The perceptual experiments (1-3) demonstrated that tonal and durational properties, which are associated with the realisation of prosodic phonological structure, are interdependent perceptual cues. Experiment 3 clearly demonstrated that tonal and durational cues play a significant role in rhythm perception. In experiment 1, native language did not affect the perceptual interdependence of duration and f_0 in non-speech stimuli or isolated syllable pairs. However, experiments 2 and 3 found that native speakers of different languages were sensitive to durational and tonal cues to different extents when perceiving rhythmic groups and judging rhythmicity, probably influenced by their native-language prosody. (Only subtle differences in the relative sensitivity to durational and tonal cues were observed between the two varieties of one language investigated here: (S)Fr.) Therefore, rhythm perception may be more language-specific than universal. Experiment 4 quantified rhythm production taking into account this acoustic multi-dimensionality and language-specificity found for rhythm perception. How these findings contribute to speech-rhythm research, by going beyond previous experiments, is explained in the following sections, which refer to the three perspectives given above.

7.1.1 A focus on perception

Despite the centuries-old observation that languages *sound* rhythmically different (e.g. Classe 1939, Roach 1982, Sievers 1912, Steele 1775), phonetic experiments on rhythm perception are vastly outnumbered by those on production. Some perceptual studies required listeners (adults or infants) to discriminate/categorise segmentally degraded stimuli from languages that were reportedly of different rhythmic 'types' (e.g. Nazzi et al. 1998, Ramus et al. 2003, Ramus and Mehler 1999, White et al. 2007). Experiment 3 in this thesis probed further into adults' perception of rhythm than these discrimination/categorisation tasks, which tell little about how acoustic properties contribute to the impression of rhythm in language that conveys meaning (cf. Barry et al. 2009), by using linguistically meaningful stimuli which make the findings generalisable to everyday speech. Moreover, this thesis did not assume the existence of a (categorical or continuous) rhythm typology. Some studies investigated perceptual isochrony, i.e. how physical and perceived timing differ (e.g. Benguerel and D'Arcy 1986, Donovan and Darwin 1979, Scott

et al. 1985). This thesis progressed beyond these studies by investigating not just perceived length, but its interaction with tonal cues in rhythm perception. Some P-centre experiments investigated how non-duration cues influenced perceptual isochrony, using monosyllabic stimuli isolated from any linguistically meaningful context (e.g. Howell 1988, Pompino-Marschall 1989, Scott 1998). This thesis (experiments 2 and 3) investigated rhythm perception in longer stretches of meaningful language.

After work began on this thesis (summer 2007), a workshop on ‘Empirical Approaches to Speech Rhythm’ was held (March 2008). Some contributions to this workshop were published in a thematic issue of *Phonetica*, which presented new directions for rhythm research that concerned production and perception of different languages and looked beyond segmental timing (Kohler 2009b). Kohler’s (2009a: 35) contribution requested the following.

‘Before physical measurement variables in speech production can be related to rhythmical patterns in a scientifically insightful way, the type and degree of rhythmicity in the data needs to be evaluated perceptually by the competent language user. Native listeners have to scale how rhythmical native speakers’ utterances are and assess whether some utterance is more rhythmical than another. It is only then that acoustic or articulatory and physiological measures can be seen as the physical exponents of rhythmic categories in speech interaction in different languages.’

In experiment 3 of this thesis, listeners did just as his second sentence suggested. The weighted PVIs (experiment 4) took, it is hoped, a first step towards finding a ‘scientifically insightful way’ to link perceived and produced rhythm ‘in different languages’.

Niebuhr (2009) and Barry et al. (2009) were also notable contributors to the *Phonetica* thematic issue. Niebuhr (2009) found that sentence-global f₀-based rhythm influenced which syllable in verbs listeners perceived as prominent (which they indicated through a speech-shadowing task). This, Niebuhr (2009) argued, supported his premise that rhythm is a perceptual phenomenon with a guide function allowing listeners to predict syllables’ perceptual properties. According to Niebuhr (2009: 109-10), rhythm is a process which is: *cyclic* (not uni-directionally hierarchical) because prominences, grouping and top-down linguistic knowledge result in global rhythm which then goes full-circle by influencing its basic constituent (locally prominent syllables); and *multi-dimensional* (not timing-based) because it involves intonation and spectral patterns. This thesis has contributed further to this conceptualisation, by demonstrating the significance of acoustic multi-dimensionality at various domains in this perceptual cycle (the syllable, the perceptual grouping of syllables, and the perceived rhythmicity of syllable-strings when top-down knowledge is relevant), and (unlike Niebuhr 2009) the *interdependence* of f₀ and duration. This suggests that a re-conceptualisation of rhythm like Niebuhr’s (2009) could lead

research forward. However, rhythm metrics might be put aside since, according to Niebuhr (2009: 95, 110), his model implies that rhythm ‘cannot be soaked up by acoustic measurements’. Indeed, Barry et al. (2009) showed that durational metrics did not capture perceived rhythms that resulted from the temporal distribution of prominences in metrical verse, and they criticised the lack of perceptual research validating the metrics. With another experiment, Barry et al. (2009) confirmed that the perceived rhythmicity of nonsense metrical verse depended on f_0 (and to some extent intensity and vowel quality) as well as duration. Their finding was made more generalisable to speech communication by the results of experiment 3 in this thesis which used meaningful sentences and obtained similar results. Experiment 4 then advanced towards a possible solution to the lack of perceptual validation of rhythm metrics; weighted (unlike durational) PVIs captured the language-specific relative significance of durational and tonal variability in perceived rhythm. We cannot claim that weighted PVIs ‘soak up’ every aspect of rhythm’s complexity, as that requires further experiments on more acoustic cues. Nevertheless, by directly linking perceived and produced rhythm, this thesis went further than all previous rhythm-metric studies, and demonstrates that acoustic measurements may be useful in rhythm research, if integrated in a perceptually informed way.

7.1.2 Inclusion of f_0

Phonetic research on rhythm has recognised that rhythm involves timing and prominence (see Adams 1979, Kohler 2009a), but timing has received most attention, and prominence has often been regarded as an abstract concept. Relatively few non-timing-based rhythm studies have appeared. Niebuhr’s (2009) and Barry et al.’s (2009) perceptual experiments were discussed above. This thesis combined investigation of interacting cues (like Barry et al. 2009) with natural-sounding meaningful speech stimuli (like Niebuhr 2009). A few production studies have quantified rhythm using a non-durational acoustic property, e.g. Low (1998) (amplitude PVI, spectral dispersion PVI), Ferragne (2008) (intensity PVI), Tilsen and Johnson (2008) (‘rhythm spectrum’: automated quantification based on Fourier analysing the amplitude envelope of bandpass-filtered speech). These metrics each incorporated only one acoustic dimension.

Nolan and Asu (2009: 68) argued that the incorporation of multiple dimensions into a rhythm metric is only profitable if we understand how those integrated dimensions interact: a matter ‘urgently in need of further research.’ Similarly, Lee and Todd (2004) concluded that rhythm experiments should investigate cue interaction and test the prediction that prominence-lending cues should display a trading relation. Lee and Todd’s (2004) rhythm metrics, based on speech output from a ‘rhythmogram’ algorithm that assigned prominence values according to duration, f_0 and intensity in the incoming signal, seemed like promising tools. This thesis addressed cue interaction, the matter needing research. Experiment 1 confirmed that one cue can influence the perception of another dimension; experiment 2 demonstrated a cross-linguistically

different ‘trading relation’ between cues to grouping; experiment 3 found that perceived rhythm results from the language-specific integration of two dimensions. These findings led to the development of rhythm metrics which incorporate how duration and f_0 interact in perceived rhythm in different languages, as Nolan and Asu (2009) suggested could be profitable. Most rhythm-metric studies have not independently justified measuring segment properties since Ramus et al.’s (1999) and Grabe and Low’s (2002) seminal papers. Experiment 4 went further than these studies by suggesting that, for language-specific metrics, it may be more appropriate to measure the acoustic variability of syllables (or larger prosodic groups) rather than vowels, because syllable variability reflects phonological structure, which native speakers have knowledge of, and which can contribute to their impression of rhythm in their language (cf. Nolan and Asu 2009).

Phoneticians who have conducted duration-based rhythm research have rarely incorporated the knowledge from other phonetic research that *various* acoustic cues can make syllables/vowels prominent. This thesis linked both areas of interest within phonetics, by investigating the consequences that interdependent cues to syllable-local prominence have for longer speech domains. Experiment 1 readdressed an existing debate over durational-tonal interdependence in syllable pairs, because the potential results had implications for duration-based rhythm research (cf. Lehiste 1976, Rosen 1977a), as turned out to be the case. Therefore, this interdependence was then investigated in longer domains. Most rhythm studies have presented ideas which developed in phonetics. P-centre research, which found that physical and perceived timing in speech differed, showed some exchange of ideas between researchers in psychology (e.g. Harsin 1997, Howell 1988, Scott 1998) and phonetics (e.g. Cooper et al. 1986, Fowler 1979, Fox and Lehiste 1987, Pompino-Marschall 1989). Yet none apparently applied their findings to modelling various languages’ rhythm, even the phoneticians, who were mostly concerned with speech perception theories. Likewise, rhythm researchers in phonetics have not generally recognised the implications of the P-centre findings for their work. This thesis originated in phonetics but considered ideas long-established in psychology: rhythm is primarily a perceived phenomenon, the investigation of which must involve more than physical duration/timing alone. The findings suggest that further inter-disciplinary research which pursues these ideas would be fruitful.

7.1.3 Cross-linguistic study

There is abundant evidence that native language influences perception of segmental contrasts (e.g. Flege et al. 1999, Hume and Johnson 2001, Lisker and Abramson 1970, Strange 1995). This thesis contributed to the less extensive evidence that native language affects prosody perception. Previous perceptual studies found that native language influenced: listeners’ ability to distinguish or recall prominence location in nonsense words (Dupoux et al. 2001, 2008); the

groupings listeners perceived in series of durationally alternating tones (Iversen et al. 2009, though not Hay and Diehl 2007); listeners' ability to learn tone sequences representing different rhythms (Bailey et al. 1999); and listeners' preferred word-segmentation strategy which may reflect their native language's rhythm (Cutler et al. 1992, Kim et al. 2008, Murty et al. 2007). Various experiments on different languages demonstrated together that the relative significance of prominence-leading cues varies cross-linguistically (see chapter 1, and studies cited by Cutler 2005: 270). This thesis started with a psycho-acoustic/-phonetic experiment like these prominence and tone-grouping/-learning experiments with non-speech, nonsense or contextually isolated linguistic stimuli. Experiments 2 and 3 then extended cross-linguistic perceptual research by investigating the implications of rhythm cue interdependence in linguistically more complex contexts, i.e. the grouping of meaningful syllables and perceived sentence rhythmicity. Unlike the segmentation-strategy experiments, this thesis examined perceived rhythm by directly eliciting listeners' intuitions, without assuming that a rhythm typology exists. Whilst experiments 2 and 3 found an effect of native language, experiment 1 did not. This is unlikely to have resulted from the difference in meaningfulness of stimuli (digits/letters and sentences mimicking communicative situations compared to words isolated from linguistic context), since previous studies with non-speech tone stimuli demonstrated some influence of native-language prosody (see above). It seems more likely that in experiment 1 the domain (monosyllable pairs) was perhaps not long enough, like multiple syllables or tones, for listeners to make some sort of 'connection/analogy' between the stimuli and prosodic structures in their native language.

Recently, some rhythm researchers have called for future investigations of cross-linguistic differences in rhythm perception which: test rhythmically different languages in identical experiments (Kohler 2009a); use linguistically meaningful stimuli (Niebuhr 2009); and examine whether cross-linguistic differences result from inter-subject variability related to rhythm's complexity or from variability dependent on subjects' native language (Barry et al. 2009). This thesis responds to all these requests. The results confirm that rhythm is complex, and they provide evidence that listeners' native language influences their perception of rhythm in speech. This supports Kohler's (2009a: 42) prediction that 'f₀, syllabic timing, syllabic energy and spectral dynamics are expected to be combined differently to create rhythmicity in these [rhythmically different] languages.'

This thesis has contributed experimental evidence which potentially supports suggestions (e.g. Dauer 1983, Roach 1982, Wenk and Wioland 1982) that listeners hear other languages' rhythms under the influence of their native prosody, which possibly contributed to the rhythm typology developed from Anglophones' impressions. Language-universal acoustically one-dimensional rhythm metrics, which have been used to support versions of that typology, are hard to justify given the results of experiment 3. Potential alternatives, suggested in experiment 4, are PVIs which language-specifically combine durational and tonal variability. The resulting

numbers less clearly distinguished two languages whose rhythms are clearly distinct according to the (categorical/continuous) typology. It is not that these languages' rhythms *are* similar, as they sound different to any given listener. Rather, if we 'cancel out' (by accounting in the metric for) the fact that the listener hears rhythm under the influence of their native language, we are left with a measure of rhythm as a language-universal phenomenon which is evidently not completely cross-linguistically divergent. This suggests that rhythm, which involves multiple interacting acoustic cues, may be universally present in languages, but what may differ between languages is how the cues are weighted and integrated relative to each other. Unlike this interpretation of weighted PVIs, Lee and Todd (2004) implied that their acoustically multi-dimensional (language-universal) metrics' weak ability to distinguish rhythm types was problematic; if their rhythmogram-derived metrics were made language-specific, it would be interesting to compare them with weighted PVIs.

7.2 Future research

Despite the contributions that this thesis has made to rhythm research, there were necessary limitations to its scope. Time and practical constraints prevented an investigation of more languages. Experiment 3 was, to my knowledge, the first to ask naïve listeners to directly judge rhythmicity in meaningful speech, so it was a pilot for future experiments, and simple stimuli were appropriate (cf. Hawkins 1999: 199 'When a complex field is only just being opened up to research, it is usually necessary to make simple assumptions and to ask simple questions.'). The same listeners might not fare so well, or even be completely baffled, if asked to judge rhythm in longer stretches of spontaneous speech. Speakers of other languages might find the task with simple stimuli impossible, perhaps e.g. Koreans (to whom prominence in their native language is not very obvious, according to Nolan and Asu 2009). Given the three perspectives advocated in §7.1, phonetic research on rhythm could benefit from taking the following next steps.

We need to conduct identical rhythm perception experiments with speakers of many diverse languages (including non-standardised dialects and regional language-varieties) which display different phonological representations and phonetic realisations of prosody (cf. Kohler 2009a, Niebuhr 2009). A few examples are: languages in which tonal patterns are (to a greater or smaller degree) lexically contrastive (e.g. Mandarin Chinese, Yorùbà, Japanese, Norwegian, see Gussenhoven 2004: chapter 3); and languages with fixed lexical stress in various positions, e.g. Finnish (initial) (Vrooman et al. 1998), Polish (penultimate) (Dogil 1999). Initial experiments should use relatively simple speech stimuli, like experiments 2 and 3, to ascertain which cues are significant in each language, and whether rhythmicity-judgement tasks are feasible for those speakers. Then stimuli should be developed which comprise longer utterances of spontaneous speech, more comparable to everyday speech communication. As chapter 5 discussed, designing cross-linguistically equivalent stimuli and tasks is difficult, but this should not deter us. Once

listeners have been tested in their native language, we could run identical tests requiring listeners to judge rhythmicity in a language they have never heard. If they find this task feasible, it would test the prediction that non-native and native listeners have different preferences for what sounds rhythmic in a language, because each group hears rhythm ‘through their own ears’ (rather than through the other group’s ears). Nevertheless, it is possible that the effect of native language on rhythm perception shown in this thesis might not be replicated with other languages. Only with a large and diverse enough sample can we draw the most robust conclusions about rhythm in ‘language’ generally.

Future research needs to link rhythm perception and production. If rhythm metrics remain in use, they need to integrate acoustic measurements in a perceptually motivated way, like the weighted PVIs. The aim of the weighted PVIs was not to ‘jump on the [rhythm-metric-experiment] bandwagon’ (Kohler 2009b: 6), but to try and steer the bandwagon around to another possible direction, which might lead to a more enlightening destination. However, weighted PVIs are not necessarily the best method for investigating rhythm, and should not, as has happened recently, dominate without us thoroughly questioning their validity in achieving the ultimate goal of better understanding rhythm. They could be a helpful tool, amongst other experimental techniques, each a means to an end, not an end in itself.

Future experiments need to investigate the interaction between several more acoustic cues (than duration and f_0) to perceived rhythm in various languages. Other obvious candidates include spectral balance (a measure of intensity distribution across the spectrum), rise time (a measure of the energy ramping in the sound-initial amplitude envelope) and formant frequencies. Sluijter and van Heuven (1996) and Sluijter et al. (1997) found that spectral balance cued perceived prominence and marked produced prominence in Dutch (for divergent results for English, see Kochanski et al. 2005). Howell (1988) and Scott (1998) found that rise time influenced English listeners’ perception of syllable isochrony. Barry et al. (2009) found that for Bulgarians, formant-frequency differences between /**ɑ**:/ and /**ə**/ were almost as significant in perceived rhythmicity as durational and tonal cues were; Low et al. (2000) found that in Singapore English, reduced vowels had F1 and F2 values more peripheral in formant-frequency space than in British English, which could be a factor in these varieties’ different-sounding rhythms.

To better understand how multiple cues are integrated in perceived rhythm, future experiments should consider the physiological and psychological bases of auditory processing widely researched in audiology (for a summary, see Moore 2004). We cannot assume that the way acoustic properties are captured visually in (and so are measurable from) a spectrogram necessarily corresponds to how those properties are captured auditorily. As an analogy, the classic London Underground map is not a visual representation of the precise physical location

of tubes underground. For a start, the spectrogram's linear frequency scale (Hertz) is unlike scales which capture how the brain perceives pitch, e.g. Bark, ERB-rate. According to Moore (2004: 51, 66), the acoustic signal entering the ear is split into component frequency bands at various points along the basilar membrane in the cochlea, so 'the auditory system behaves as if it contains a bank of bandpass filters, with overlapping passbands.' Sluijter and van Heuven (1996) considered this in their experiment on prominence. Likewise Harsin's (1997) P-centre model (like, he noted, Scott and Howell's) considered amplitude-modulation rates within different frequency bands which (audiology research had found) related to perceived speech rhythms. (However, Villing et al. (2003) argued that these P-centre models did not adequately approximate the cochlear filtering process). How the speech signal is transformed in the auditory nerve and 'understood' in the brain is not fully understood (Hawkins 1999: 216); recent developments in neuroimaging may shed light on this (see Pulvermüller 2002: 44-49). Rhythm research using neuroimaging techniques (e.g. Jomori and Hoshiyama 2009) should increase our understanding of how multiple rhythm cues are integrated with each other during speech processing in the human brain, including the cognitive influence of a listener's particular linguistic system.

7.3 Consequences of this research

The findings that this thesis (§7.1) and future experiments (§7.2) contribute to rhythm research have implications for theoretical models and for practical applications.

7.3.1 Implications for theoretical models

7.3.1.1 Modelling cross-linguistic variation in rhythm

The 'stress-timing/syllable-timing' theory dates back several decades (e.g. Pike 1945), and was made into a rhythm typology by Abercrombie (1967). This typology, though widely criticised, still pervades discussion on rhythm in phonetics, particularly amongst rhythm-metric studies. However, many researchers' ideas could be seen as general descriptions of cross-linguistic rhythmic variation, and are no longer claims of a categorical typology (cf. Cummins 2002). It now seems time (no pun intended) to develop a new model to explain cross-linguistically variable rhythm, not grounded in remnants of a timing-based theory. Although rhythm *involves* timing, rhythm *does not equal* timing. Rhythm involves a complex interaction of multiple acoustic properties of prominent and non-prominent sections of speech. It is possible that speakers of a language subconsciously agree to exploit some acoustic properties more than others (depending on several factors like how each property realises aspects of the language's phonology), to produce a rhythm that, when perceived by other speakers of that language, aids communication between them. Languages vary in phonology and phonetic realisation, so the most exploited properties may differ cross-linguistically, i.e. the attention speakers give to certain acoustic cues depends on their native language.

Figure 7-1 is not a comprehensive model of speech rhythm. Rather it illustrates how future rhythm research should differ from duration-based research, which compares between languages just one part of this diagram. If we, for now, conceptualise rhythm as a perceptual phenomenon which results from the multi-dimensional acoustic signal and the influence of listeners' native language, future research should compare between languages *all* of this diagram. We can refine this outline of rhythm in light of these future experiments' results.

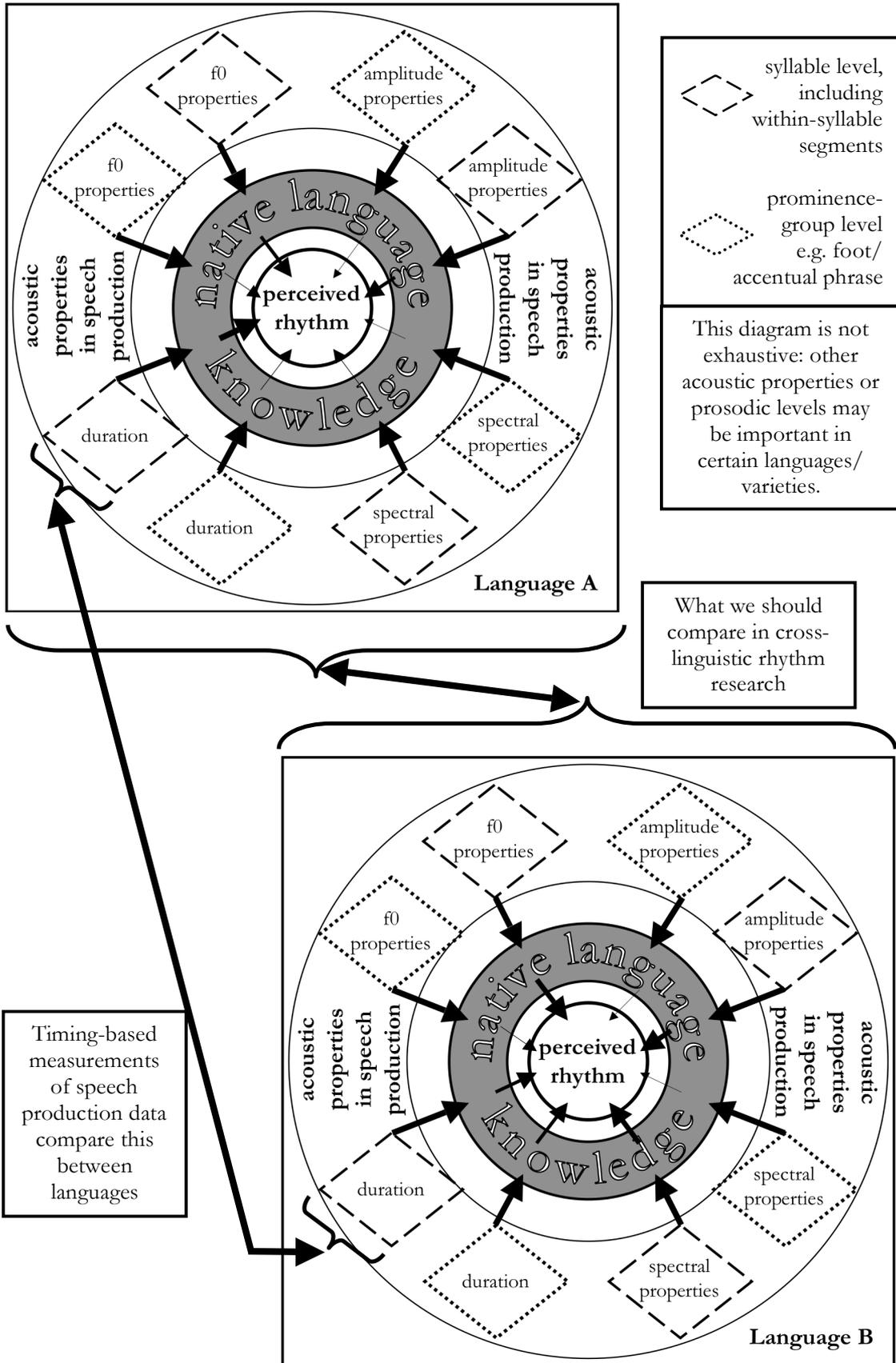


Figure 7-1 – Cross-linguistic comparisons of durational measurements do not compare every factor that contributes to (represented by arrows) perceived rhythm; listeners’ native language (which the arrows pass through) influences which cues are more important (thicker arrows) than others (thinner arrows).

7.3.1.2 Speech perception models

As rhythm is a perceptual phenomenon, speech perception models should account for (current and future) empirical data on the multi-dimensionality and language-specificity of rhythm. Current theories of speech perception are based on investigation biased towards segmental perception (cf. Hawkins and Smith 2001: 3-5, House 1990: 9, and references therein), with relatively far less known about prosody perception (cf. Vaissière 2005). According to Hawkins and Smith (2001: 34), '[n]o current speech understanding models can adequately explain the percept of rhythmic structure.' House (1990) proposed a pitch perception model, based on empirical data, which postulated that tonal-movement perception is constrained by rapidly changing spectral configurations. According to House (1990: 142), one implication of this model 'could be separate perceptual mechanisms for segmental and tonal cues', with segmental mechanisms favouring rapid spectral changes and discontinuities, and tonal mechanisms favouring spectral stability for optimal perception of tonal dynamicity. Related to House's (1990) findings for pitch, there is some evidence that the perceptual salience of amplitude is associated with its dynamicity: Harsin (1997) (and others he cited) found that peaks in amplitude rate-of-change correlated with P-centres.

Like the argument of this thesis that rhythm perception depends on language-specific cue weighting, House (1990: 143) argued that although his model was theoretically applicable to any language, the details of the spectral constraints on tonal perception may vary cross-linguistically. Table 7-2 names some speech perception models that consider how native language might constrain perception.

Model	Based on data from...	Postulates that...
Native Language Magnet model	infant language acquisition	'exposure to language early in life produces a change in perceived distances in the acoustic space underlying phonetic distinctions, and this subsequently alters both the perception of spoken language and its production.' (Kuhl and Iverson 1995: 122)
Perceptual Assimilation Model	native-language (L1) influence on adults' perception of a language unknown to them	'non-native segments [...] tend to be perceived according to their similarities to, and discrepancies from, the native segmental constellations that are in closest proximity to them in native phonological space.' (Best 1995: 193)
Speech Learning Model	L1 influence on adults' perception of a second language (L2) they speak	language-specific aspects of speech sounds are stored in L1 and L2 phonetic categories which are related in a common phonological space; the SLM predicts that an L2 category may not be formed for a sound phonetically similar to an L1 sound, if the L1 phonology filters out phonetically important properties of the L2 sound. (Flege 1995: 238)

Table 7-2 – Models that consider how native language might constrain perception

These models propose different perceptual primitives and mechanisms, and their research perspectives differ, hence the types of listeners who provided the data. Yet they offer similar explanations for how listeners' L1 constrains perception. That is, essentially, that babies are born with an abstract phonological or acoustic 'space' in the brain which becomes 'set' in some language-specific way as they hear an ambient language in infancy, and this L1 'setting' is the reference against which they compare all other languages as they hear them throughout life. These models only describe this 'space' in terms of segment-sized units, and generally exemplify with data from consonant or vowel discrimination/categorisation experiments. However, such models might be extendable to rhythm perception, if researchers were motivated that way.

Conversely, rhythm perception is a 'clearly crucial' part of the *PolySp* (POLYsystemic SPEech) perception model (Hawkins and Smith 2001: 18). According to Hawkins and Smith (2001), other speech perception models are wrong to axiomatically assume the phoneme's primacy and to seek perceptual correlates for phonemes. In *PolySp*, linguistic units' perceptual correlates are complex, distributed over relatively long domains (e.g. syllable, rhythmic group, IP), and contribute to several units simultaneously (Hawkins and Smith 2001: 5), so segmental and prosody perception are considered together. *PolySp*, like other models, postulates that cognitive phonetic categories are ambient-language-specific, since they emerge from speech input and pre-existing relevant knowledge; unlike in other models, these categories are multi-modal, dynamic, context-sensitive and labile (see Hawkins and Smith 2001: 23-30). Although rhythm is evidently regarded as primarily a temporal phenomenon in *PolySp*, other cues need not be excluded, since this model postulates that temporal *and* spectral information in the signal is important in perceiving speech generally. Furthermore, Hawkins and Smith (2001: 35) argued that 'the percept of rhythm in speech is at least partly language-specific.' *PolySp* can account for the findings of this thesis that perceived rhythm is multi-dimensional and language-specific. For example, infants hearing SG spoken around them might form categories specifying that it is more important to attend to pitch than length cues for perceiving rhythmic-group boundaries, but more important to attend to length than pitch cues for perceiving rhythmical patterns. (Other acoustic cues might also be important.)

Ultimately further research needs to test *how* we come to 'listen out' for cues that matter and 'filter out' superfluous acoustic information, when perceiving rhythm in our native language. In any language, the primary rhythm cue might interact with less important cues, as in a trading-relations situation, which is well-attested for segmental contrasts and was seen for rhythmic grouping in experiment 2. This further research could clarify whether a rhythm perception model should have mechanisms associated with rhythm cues which are separate from segmental perception mechanisms (as House 1990 suggested), or whether the model should not distinguish prosodic from segmental perception (as Hawkins and Smith 2001 proposed). Neuroimaging studies need to investigate how perceptual mechanisms (i.e. a model's categories/constraints)

associated with rhythm might be structured physiologically in terms of neurons in the brain (cf. Hawkins and Smith 2001). As we understand more about rhythm perception, models of cross-linguistic rhythm variation and speech perception models can complement each other and improve.

7.3.2 Implications for practical applications

Beyond theoretical modelling, the practical applications which could benefit from research on the multi-dimensionality and language-specificity of rhythm include speech technology, L2 teaching, and remediation of developmental language disorders. These are now discussed in turn.

7.3.2.1 Speech technology

Several publications have identified that prosody research is important in developing automatic speech recognition (ASR) systems that recognise spontaneous speech, and speech synthesis systems that generate more natural-sounding speech (e.g. Barry et al. 2005: 2, Holmes and Holmes 2001: 247, Keller et al. 2002: 87-196). Very few ASR systems have incorporated prosody (Batliner and Möbius 2005, Holmes and Holmes 2001). Batliner and Möbius (2005) outlined a prosodic model for use in ASR, similar to those of the few others working on this, which incorporated all available information on multiple acoustic properties (f_0 , energy and duration) plus other linguistic information. If ASR systems are to incorporate prosody like this, data on the language-specific integration of multiple rhythm cues (from this thesis and future research) would be useful.

Conversely, there has been extensive incorporation of prosody in speech synthesis systems, as it is deemed necessary for human-like speech (Batliner and Möbius 2005, Zellner Keller 2002). Most systems compute prosodic timing (usually syllable durations) first, and then intonation (i.e. an f_0 contour) (Zellner Keller 2002), and some include intensity modifications (Holmes and Holmes 2001, Monaghan 2002). However, Batliner and Möbius (2005), citing House (1990), suggested that this synthesis of properties separately is inconsistent with evidence that speakers co-produce tonal, temporal and spectral properties to allow listeners to optimally perceive prosody. Similarly, van Santen (2005: 163) suggested that speech synthesis could be improved by ‘multidimensional modeling of all acoustic prosodic features – F_0 , local acceleration, spectral balance, loudness, etc.’, and by research on how durational and spectral properties could be *integrated* to mimic natural prosody. Batliner and Möbius (2005: 36) stated that synthesis systems which have ‘gone multilingual’, by including a prosody/intonation module with both language-independent features and language-specific input, require more research on what is language-independent/-specific in prosody. Experimental phonetic research on the multi-dimensionality and language-specificity of rhythm could help in the development of prosodic

models for synthesis which integrate multiple rhythm cues simultaneously and language-specifically.

Barry et al. (2005), Batliner and Möbius (2005) and van Santen (2005) were all contributors to a conference session about the mutually beneficial interdisciplinary communication between phonetics and speech technology. Phoneticians can use speech synthesis as a tool for investigating rhythm (in various languages), by observing the system's prosody settings when it generates its most natural-sounding speech. This is what Zellner Keller (2002) did for French rhythm using a speech synthesis system which modelled temporal properties at several interacting levels (segment, syllable, phrase). Siebenhaar et al. (2004) argued that their use of speech synthesis for investigating prosody forced them to model *all* prosodic aspects together, unlike traditional analytic methods which generally compartmentalise research into e.g. 'rhythm' or 'intonation'. (However, speech technologist Monaghan (2002) argued that the complexity of prosody meant improvements in speech synthesis would come from separate experiments each on one prosodic aspect.) Speech synthesis systems can, therefore, test how segmental and prosodic information are interrelated, how multiple prosodic cues interact, and how this differs between languages/dialects/varieties (Siebenhaar et al. 2004). These questions are like those which this thesis posed and advocated for future research.

7.3.2.2 L2 teaching

According to Trouvain and Gut's (2007) preface to the proceedings of a workshop on teaching L2 prosody, interdisciplinary communication between (theoretical) phonetics and (applied) L2 pedagogy would be mutually beneficial. In phonetics, research on L2 speech mostly concerns segments, rarely prosody (according to: Boula de Mareüil and Vieru-Dimulescu 2006, de Bot 1986, Munro and Derwing 1999, Piske et al. 2001, Trouvain and Gut 2007). Likewise in L2 pedagogy, teaching materials and publications on instruction rarely concern prosody or pronunciation in general (according to: Derwing and Munro 2009, Hirschfeld and Trouvain 2007, Mehlhorn 2007, Trouvain and Gut 2007). Specifically for rhythm, Wenk (1986: 120) pointed out the 'dearth of materials available to learners on the rhythms of speech.' However, speakers' realisation of their L1 prosodic properties when producing L2 has been found to contribute considerably to their non-native accent and/or intelligibility¹ (e.g. Boula de Mareüil and Vieru-Dimulescu 2006, Munro 1995, Munro and Derwing 1999, Thorén 2008: 122-4, White and Mattys 2007b). Together these experiments demonstrated that L1 prosodic influence in non-native accents involves multiple acoustic properties like f_0 , duration and their interaction, and that the relative significance of each cue in how foreign L2 speakers sound might vary between

¹ Perceived strength of foreign accent does not necessarily correlate with measured intelligibility (see Derwing and Munro 2009).

target languages. Therefore, if L2 learners' aims are to avoid a marked foreign accent (which may attract native speakers' negative attitudes, see e.g. Derwing and Munro 2009), and to be intelligible enough for communication with native speakers, L2 teaching should include prosody and consider which prosodic cues are most important in the target language. For this purpose, data on the language-specific integration of various rhythm cues would be useful.

Currently, teaching materials which do include rhythm have a too simplistic notion of regularly timed stress or syllable 'beats', according to Wenk (1986) and Barry (2007) who also critiqued the simplistic stress-timing/syllable-timing typology in theoretical phonetic publications (e.g. Wenk and Wioland 1982, Barry et al. 2003, 2009). The question is how rhythm teaching materials might be improved to recognise rhythm's multi-dimensionality and language-specificity, which makes a simple language-universally applicable method unfeasible (cf. Barry 2007). It seems L2 learners would benefit from somehow recognising the significance of certain acoustic cues over others in the L2. This probably (depending on the L1-L2 combination) means that learners need to rearrange the cues' significance compared to in their L1, as if 'hearing through the ears' of a native speaker of their L2. This may be difficult, since several acoustic cues are involved and may contribute to rhythm in only subtly different ways cross-linguistically. The following methods, which involve drawing learners' attention to various acoustic properties of segments, prosodic groups or utterances, are possibilities: teaching learners how to (not) reduce vowels or differentiate long and short vowels in production (Barry 2007); making learners self-monitor their L2 pronunciation in terms of intonation and prominence ('Contrastive Prosody Method', Missaglia 2007); making learners indicate boundaries when hearing lexically identical utterances pronounced with differently patterned tonal and durational properties (Hirschfeld and Trouvain 2007); allowing learners to visualise in *Praat* differences in intonation, syllable durations and intensity between their L2 productions and equivalent native-speaker productions (Mehlhorn 2007). Further research to test how learners could be helped to learn the target-language complex rhythmic-cue structure can be complemented by theoretical phonetic experiments on the multi-dimensionality and language-specificity of rhythm.

It may be an onerous task to develop different language-specific teaching methods for rhythm. Yet it seems worthwhile given some evidence from real teaching environments: learners who received prosody-focussed training were judged by teaching experts to be more comprehensible and to have better pronunciation than those who received segment-focussed training (Derwing et al. 1998, Missaglia 2007). (See, though, Barry's (2007: 114) suggestion that a focus on individual segmental/syllabic problems is more feasible for improving learners' rhythm than an all-in-one 'rhythmic blanket' method.) Nevertheless, given Galloway's (2007) findings for French~Swiss-German bilinguals who learned L2 from adolescence onwards in a naturalistic non-educational environment, learners need not *necessarily* become 'meta-linguistically' aware of the target rhythm through formal instruction to master native-like L2 rhythm (as measured with

durational PVI). However, these bilinguals' rhythm productions might not be as native-like if measured with weighted PVIs, which account for multiple language-specifically weighted cues.

7.3.2.3 Remediation of disorders in L1 acquisition

Unlike in L2 learning through education, infants generally grasp unproblematically how rhythm cues are weighted in their ambient (to-be-native) language. Yet Goswami and colleagues have found that children and adults with developmental language disorders (dyslexia or specific language impairment, SLI) have problems auditorily processing some rhythm cues, particularly rate of amplitude-envelope onset i.e. rise time (e.g. Corriveau and Goswami 2009, Corriveau et al. 2007, Goswami et al. 2009, Hämäläinen et al. 2009, Pasquini et al. 2007). The explanation that these studies give is as follows (see e.g. Corriveau et al. 2007: 663-64, Goswami et al. 2009: 20-21). Rhythm cues allow infants to identify the prominence patterns of their ambient language and thus to develop a suitable word-segmentation strategy (e.g. Ramus, Mehler, Jusczyk and their colleagues' findings). If certain infants are less (or in)sensitive to these cues, this could impair their identification of rhythmic patterns and hence their segmentation of syllables and words, leading to the development of degraded within-syllable phonological representations, which could then disorder their literacy acquisition (i.e. dyslexia). Moreover, prominence patterns may relate to syntactic constructions in the ambient language, so insensitivity to rhythm cues could impair infants' ability to acquire distinctions in syntax and other related areas of their language system (i.e. SLI).

These experiments tested rhythm-cue processing deficits using non-speech stimuli. Hämäläinen et al. (2009) found that Finnish dyslexic children were more sensitive to rise-time manipulations than English dyslexic children tested with identical stimuli by Richardson et al. (2004), relative to the respective control subjects, who were more sensitive than both dyslexic groups. Therefore, rhythm-cue processing impairments in dyslexia may be language-universal with subtle language-specific differences, but research with more languages is needed on this (cf. Corriveau et al. 2007, Hämäläinen et al. 2009). Data on the language-specific integration of multiple rhythm cues is thus needed for language-disordered listeners too. We could run experiments as in this thesis with dyslexics and SLI sufferers, to test the prediction that language-disordered adults might be less sensitive to the cues favoured by normally developed speakers of the same language when locating rhythmic-group boundaries and judging sentence rhythmicity.

Ultimately, research on the multi-dimensionality and language-specificity of rhythm in normal and impaired listeners can help develop language-disorder remediation methods. Young dyslexics and SLI sufferers could be helped to attend more to the cue(s) which is/are found to be most important in their language. Goswami (personal communication) and colleagues intend to develop a rhythm-based remediation for young SLI sufferers which they have trialled (with promising results) with young dyslexics. In both cases the L1 is English; activities like 'playing a

drum in time with the stressed syllables in nursery rhymes (HUMP-ty DUMP-ty SAT on a WALL)' (Corriveau and Goswami 2009: 129) would not be applicable to other languages, e.g. Korean. Further research could lead to other methods suitable for different languages.

7.4 Does rhythm exist in speech?

The answer to this important question could be, at least for some languages, 'no', according to Pamies Bertrán (1999: 127) and Nolan and Asu (2009: 76). If so, this would seem a little disappointing for a thesis entitled 'Speech rhythm: [...]', though it would not necessarily be an anticlimax (in any research project) to conclude something that few predecessors had even questioned at all. However, in this thesis, naïve listeners were able to judge sentence rhythmicity without being given any definition of rhythm (chapter 5), which suggests that some regular pattern was perceptible in the (SG or Fr) signal. It might be argued that listeners could have based their judgements on timing (i.e. whether the duration of all segments/syllables was appropriate), which would be possible even if the stimulus were arrhythmic (i.e. had no regular pattern). As an analogy, imagine a piece of cloth (X) with random colour patches along its length (to represent arrhythmicity), and two other similar pieces (A, B) with the same order of colour patches as in X, but some patches lengthened to different extents in A and B; we could probably say whether A or B better matched X, a judgement based purely on the length of patches, as these had no regular pattern. However, according to what these naïve listeners said contributed to their impression of a regular pattern in the signal, most did not base their judgements purely on timing (see Figure 5-18). For (S)Fr, intonation, accentuation and speed/timing were each given by a similar number of listeners as factors which contributed to their impression of rhythm; for SG, stress/accentuation was given by more listeners than speed/timing or length of vowels/syllables (several said they relied on intuition so did not indicate which particular cues contributed).

Rhythm involves multiple acoustic properties of speech (and these naïve listeners noticed this, given their impressions just summarised). Although these listeners recognised rhythm in speech, if the Fr listeners were to listen out for certain patterns in intonation, accentuation or speed in SG speech, or if the SGs were to listen out for certain patterns in stress or vowel length in Fr speech, they would not hear such patterns. Rhythm in a particular language also involves the relative significance of the multiple acoustic properties as perceptual cues. The results of experiment 3, which built on evidence from experiments 1 and 2, do not undermine the working definition of speech rhythm given in chapter 1 as:

- (1) the **perceived** regularity in an utterance,
- (2) induced by the acoustic **multi-dimensionality** of the speech signal (which results from a realisation of phonological structure),

(3) and influenced by the listener's **native language**.

In the experiments, the focus on **perceived** rhythm, the inclusion of **more than one** acoustic cue, and the testing of **native-language** influence in perception have all proven worthwhile. The language-specific integration of constantly varying interdependent acoustic cues makes speech rhythm more complex than, say, a simple drum beat; a drumstick repeatedly hitting a drum creates a signal varying little in resonance frequencies or f_0 , though the amplitude and timing of the hitting sound may vary. Similarly, for music, Pamiès Bertran (1999: 126) remarked that although music often has an underlying percussive beat, its rhythm is not necessarily simplistic, as several phenomena can occur within the complex signal which complicate the timing beyond the beat.

The above 'working' definition was designed to be testable and open to reformulation. Although the evidence from this thesis supports the argument that rhythm exists in speech, future research should investigate many more languages and use longer spontaneous speech stimuli (see §7.2). Nolan and Asu (2009: 76) stated that '[i]t is always possible that we have been misled into overestimating the degree of rhythmicity in speech by a subset of languages which provide prominences which, given our tendency to seek patterns in what we perceive, allow us to imagine a rhythmicity which is not there.' Vaissière (1991a: 259, 1991b: 109) stated that Fr, unlike English, lacks a recurring strong prominence, but this actually makes Fr rhythm 'much more obvious' to her (a native Fr speaker) than English rhythm. Given that in this thesis (S)Fr listeners perceived rhythm in speech, despite this language not being in the traditional 'subset [...] which provide prominences', we may not have been totally misled. Yet it is possible that any listener seeks patterns and imagines rhythmicity when their attention is focussed on the very task of perceiving something called rhythm, i.e. experts listening to languages with that intention, or naïve listeners in rhythmicity experiments. Thus an experimental paradigm which taps into listeners' intuitions about speech rhythm without drawing their attention to it would be desirable, though this seems paradoxical.

To summarise an answer to the question 'does rhythm exist in speech?', there is for now enough evidence that rhythm exists in speech to make it worthwhile continuing investigation of the phenomenon (as proposed in §7.2), but we must always consider that future results could contradict this conclusion. Although rhythm is not completely non-existent in speech, it is certainly not obvious in the complex acoustic signal, and what is obvious to a speaker of language X may not be obvious to a speaker of language Y. In all languages, multiple acoustic cues may integrate and contribute to perceived rhythm in speech; what may differ between languages is how much each cue contributes relative to the others.

Much recent research on rhythm comes from the perspective that each language 'has' a rhythm, which is measurable from the acoustic signal produced by speakers, and which pre-

lingual infants can distinguish from the rhythm of other languages. Yet the perspective that ‘rhythm is in the ear of the beholder’ (cf. Kuhl and Iverson 1995) may be more appropriate, given that this thesis’ definition of rhythm implies that rhythm in a language, say X, is only available for native speakers of X to perceive. Infants can extract a regular pattern from the acoustic signal of X and distinguish it from the regular pattern in another language, as can adults who have never heard X before, and these adults may perceive the regular pattern in X by comparison with their native-language phonology. However, in neither the infant nor the naïve adult case is the perceived rhythm complete: native-speaker knowledge about which cues are more important than others needs to accompany the perceptible regularity.

Appendices

The following sections present, for each experiment, the instructions that subjects read before the perceptual tasks (chapters 3-5), and results tables with values which were displayed only in graphs in chapters 3-6. All the instructions are English translations of the (standard) German and Fr versions that I wrote for SG and (S)Fr subjects respectively. Also presented are the texts for experiment 4 (chapter 6), and the stimulus sentences and post-test questionnaire (in English translation) for experiment 3 (chapter 5).

8.1 Experiment 1 (chapter 3)

8.1.1 Instructions

These instructions were read by the subjects who heard the buzz stimuli first. Those who heard the [si] stimuli first read instructions almost identical to these, but ‘Part 1’ and ‘Part 2’ were in the reverse order and the summary was adjusted accordingly.

Instructions

The experiment has two parts that are very similar.

Part 1

In this part you will hear many pairs of buzzes. For each pair, your task is to decide which buzz was longer.

If the first buzz was longer, click the mouse on the yellow rectangle labelled ‘1 was longer’. If the second buzz was longer, click the mouse on the yellow rectangle labelled ‘2 was longer’.

Some pairs will be much easier than others. Please do your best to decide which buzz was longer, even if they sound very similar in length. You will have as much time as you want to decide after you have heard each pair, but don’t spend too long deciding – just respond according to your initial thought.

There will be a practice session before the main experiment – this will be indicated on screen.

Part 2

In this part you will hear many pairs of sounds which sound like the word ‘si’ [if/yes, she/they]. For each pair, your task is to decide which ‘si’ was longer.

If the first ‘si’ was longer, click the mouse on the yellow rectangle labelled ‘1 was longer’. If the second ‘si’ was longer, click the mouse on the yellow rectangle labelled ‘2 was longer’ (you will get the idea by now after having completed part 1).

Again, some pairs will be much easier than others. Please do your best to decide which ‘si’ was longer, even if they sound very similar in length. You will have as much time as you want to decide after you have heard each pair, but don’t spend too long deciding – just respond according to your initial thought.

Again, there will be a practice session before the main experiment – this will be indicated on screen.

Summary

In both parts you must:

- decide which buzz or ‘si’ is longer in each pair.
- take a break regularly – this will be indicated on screen.
- complete the practice session – please follow the instructions on screen.

If you have any questions now, or after the first practice session, please ask me.

8.1.2 Results

Language		Buzzes			[si] stimuli		
		'F>L'	'R>L'	'C>L'	'F>L'	'R>L'	'C>L'
SG (<i>k</i> =28)	\bar{x}	66.27	63.69	62.95	69.25	63.29	71.28
	<i>s</i>	19.65	16.73	17.36	14.46	15.07	16.05
SFr (<i>k</i> =30)	\bar{x}	62.78	67.59	61.39	74.26	68.89	76.67
	<i>s</i>	18.92	16.32	19.08	11.44	11.07	14.33
Fr (<i>k</i> =28)	\bar{x}	71.63	69.44	69.35	72.42	68.06	73.36
	<i>s</i>	20.47	19.63	22.32	17.53	16.12	16.49

Table 8-1 – Data from Figure 3-3. Mean and standard deviation (across *k* subjects) of the percentages of ‘falling/rising/complex is longer than level’ responses: ‘F>L’, ‘R>L’, ‘C>L’; percentages out of 18, 18 and 24 trials for F, R and C stimuli respectively.

		Buzzes				[si] stimuli			
		'LF>L'	'EF>L'	'LR>L'	'ER>L'	'LF>L'	'EF>L'	'LR>L'	'ER>L'
All languages (<i>k</i> =86)	\bar{x}	70.85	63.03	66.91	69.30	69.84	75.30	69.74	64.76
	<i>s</i>	23.28	23.41	22.97	23.94	20.94	18.34	20.58	20.41

Table 8-2 – Data from Figure 3-4. Mean and standard deviation (across *k* subjects) of the percentages of ‘late fall/early fall/late rise/early rise is longer than level’ responses: ‘LF>L’, ‘EF>L’, ‘LR>L’, ‘ER>L’; percentages out of 6 trials each for LF, EF, LR and ER stimuli.

Language		Buzzes			[si] stimuli		
		'250D>L'	'375D>L'	'500D>L'	'250D>L'	'375D>L'	'500D>L'
SG (<i>k</i> =28)	\bar{x}	61.96	66.96	63.57	58.39	73.39	73.04
	<i>s</i>	21.27	16.91	17.26	16.22	15.81	17.29
SFr (<i>k</i> =30)	\bar{x}	61.00	66.67	63.33	67.50	76.83	76.33
	<i>s</i>	20.57	16.73	21.23	15.63	14.83	11.06
Fr (<i>k</i> =28)	\bar{x}	64.64	73.21	72.32	62.32	76.96	75.36
	<i>s</i>	20.72	21.82	21.96	17.66	18.87	15.51

Table 8-3 – Data from Figure 3-5. Mean and standard deviation (across *k* subjects) of the percentages of ‘dynamic is longer than level at 250ms/375ms/500ms’ responses: ‘250D>L’, ‘375D>L’, ‘500D>L’; percentages out of 20 trials each for 250ms, 375ms and 500ms stimuli.

Language		Buzzes		[si] stimuli	
		'D1>L'	'D2>L'	'D1>L'	'D2>L'
SG (<i>k</i> =28)	\bar{x}	66.55	61.79	68.21	68.33
	<i>s</i>	19.36	19.08	17.25	15.62
SFr (<i>k</i> =30)	\bar{x}	69.67	57.67	76.22	70.89
	<i>s</i>	20.22	17.92	11.93	12.10
Fr (<i>k</i> =28)	\bar{x}	74.17	65.95	74.29	68.81
	<i>s</i>	22.42	21.36	16.65	18.10

Table 8-4 – Data from Figure 3-6. Mean and standard deviation (across *k* subjects) of the percentages of ‘dynamic is longer than level when dynamic is first/second’ responses: ‘D1>L’, ‘D2>L’; percentages out of 30 trials each for D1 and D2 stimuli.

8.2 Experiment 2 (chapter 4)

8.2.1 Instructions

These instructions were read by the subjects who saw the 3+2 grouping on the left of the screen. Those who saw the 3+2 grouping on the right of the screen read instructions almost identical to these, but the visual examples illustrated the 3+2 grouping on the right.

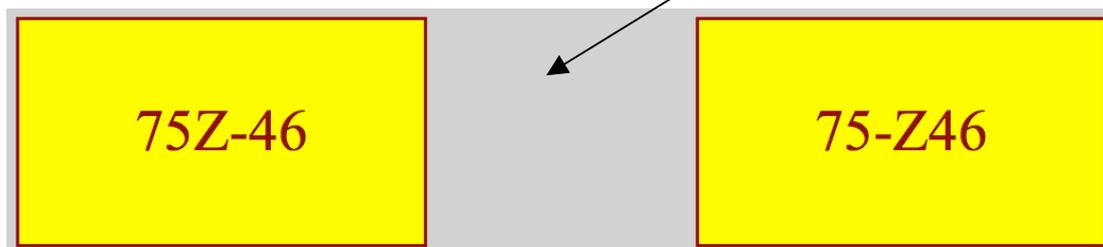
Thank you for taking part in this listening study about your comprehension of numbers and letters.

Instructions

You are going to hear a man speaking several sequences of five numbers or letters. Some sequences contain numbers and letters, some contain just numbers, some contain just letters. Some will sound more natural than others, but don't pay too much attention to the pronunciation, as that is not the object of the exercise.

For each sequence that you hear, indicate whether you think it is best grouped as either 'XXX – XX' or 'XX – XXX'. For example, in the following sequence you would hear '7 5 Z 4 6', and you would need to decide:

would this sequence be grouped as '75Z – 46' or '75 – Z46'?



Click the mouse on the appropriate yellow rectangle to give your answer. Don't think too much about your response; I am more interested in your intuitive reaction. Some sequences may be more obvious than others, but please choose the option that is the BEST representation (in terms of groups) of what you heard. After a short pause, you will hear the next sequence.

You will have the opportunity to take a short break five times during the experiment (this will be indicated on screen). There is a short practice session before the main experiment.

Summary

- Listen carefully to each sequence of five numbers/letters.
- Decide whether it is best grouped as 'XXX – XX' or 'XX – XXX', and click the appropriate yellow rectangle.
- Take a short break whenever instructed on screen.

If you have any questions now, or after the practice session, please ask me.

8.2.2 Results

Language		Digit/letter pattern					
		BBBBB	22222	3PTP3	5Z4J6	H2H2H	S888F
		PPPPP	33333	2BDB2	75Z46	C3C3C	S888F
SG	\bar{x}	69.44	70.63	51.98	42.86	44.84	62.30
($k=36$)	s	29.60	35.64	35.18	28.57	31.33	35.51
SFr	\bar{x}	75.56	72.18	59.77	47.74	56.39	56.39
($k=38$)	s	19.90	20.99	23.22	23.08	26.77	26.35
Fr	\bar{x}	67.86	69.44	53.57	37.70	45.63	49.21
($k=36$)	s	26.75	23.94	31.37	33.31	27.71	29.46
BiSG	\bar{x}	67.86	73.57	42.86	41.43	42.14	58.57
($k=20$)	s	25.75	23.30	24.53	20.15	30.21	23.59
BiSFr	\bar{x}	70.71	68.57	55.71	46.43	41.43	57.14
($k=20$)	s	24.73	20.52	18.48	21.68	24.48	23.17

Table 8-5 – Data from §4.6.3: mean and standard deviation (across k subjects) of percentage difference scores (averaged across all cue conditions)

Language		Cue condition						
		length	pitch1	pitch2	pitch1 & length acc	pitch2 & length acc	pitch1 & length con	pitch2 & length con
SG	\bar{x}	66.20	49.07	71.30	76.39	83.80	43.52	8.80
($k=36$)	s	35.96	28.71	37.50	37.03	35.74	37.85	42.44
SFr	\bar{x}	73.68	40.79	59.65	85.09	87.28	52.63	30.26
($k=38$)	s	22.80	27.86	27.30	23.50	22.41	35.41	39.68
Fr	\bar{x}	66.67	32.87	52.78	74.07	80.56	43.52	26.85
($k=36$)	s	30.34	37.69	34.39	35.06	29.68	35.02	42.59
BiSG	\bar{x}	60.83	40.83	67.50	82.50	86.67	33.33	9.17
($k=20$)	s	24.94	26.75	33.54	20.57	22.03	28.10	35.65
BiSFr	\bar{x}	71.67	39.17	62.50	86.67	88.33	38.33	10.00
($k=20$)	s	25.42	30.72	29.06	21.36	18.02	34.67	36.03

Table 8-6 – Data from Figure 4-7: mean and standard deviation (across k subjects) of percentage difference scores (averaged across all digit/letter patterns)

8.3 Experiment 3 (chapter 5)

8.3.1 Stimulus sentences

For each of thirty sentence pairs, the SG sentence appears above the Fr sentence, with an English translation of each language.

Sentence	Translation
Min Vatter isch nonig go schaffe Mon grand-père a perdu son emploi	My father hasn't gone to work yet My grandfather has lost his job
DStudäntin hät gaar nüt verstande L'étudiante a compris les questions	The student understood absolutely nothing The student understood the questions
En Gèèrtner söll dBirke go schniide L'agronome doit couper le bouleau	A gardener should cut the birch tree The farmer must cut the silver birch tree
En Tiger hät dSchòòffli verschlunge Une tigresse a mangé l'agneau noir	A tiger ate the lamb A tigress ate the black lamb
E Fläsche chan sehr schnell verbräche Une bouteille s'est cassée facilement	A bottle can break very quickly A bottle broke easily
Sin Naachber wiirt dBuebe go hole Son voisin va chercher les garçons	His neighbour will fetch the boys His neighbour will fetch the boys
Sin Unkel isch wider go bade Son beau-père s'est baigné dans le fleuve	His uncle went swimming again His step-father swam in the river
De Michi wott sHanni vergässe Nicolas veut quitter sa copine	Michi wants to forget Hanni Nicolas wants to leave his girlfriend
En Becker mues dWeggli go bache Un boucher doit rôtir du poulet	A baker must bake the bread rolls A butcher must roast some chicken
DFranziska isch naime go tschogge Marie-Claire avait fait du jogging	Franziska has never been jogging Marie-Claire had been jogging
De Leerer wiirt dSchüeler erchäne L'enseignante a connu les élèves	The teacher will recognise the pupils The teacher recognised the pupils
De Röbi mues sTrinkgält verdiene Paul-Henri doit gagner de l'argent	Röbi must earn some tips Paul-Henri must earn some money
Morn Morge mues dLeila a dUni Demain soir Laure va voir ce bon film	Tomorrow morning Leila must go to uni Tomorrow evening Laure is going to see a good film
Gescht Abig händ dStudis zviel trunke Hier matin plein de mecs ont trop bu	Yesterday evening the students drank too much Yesterday morning lots of guys drank too much
De Reto hät dRosi betroge Jean-Christophe a trompé la jeune Rose	Reto deceived Rosi Jean-Christophe deceived young Rose

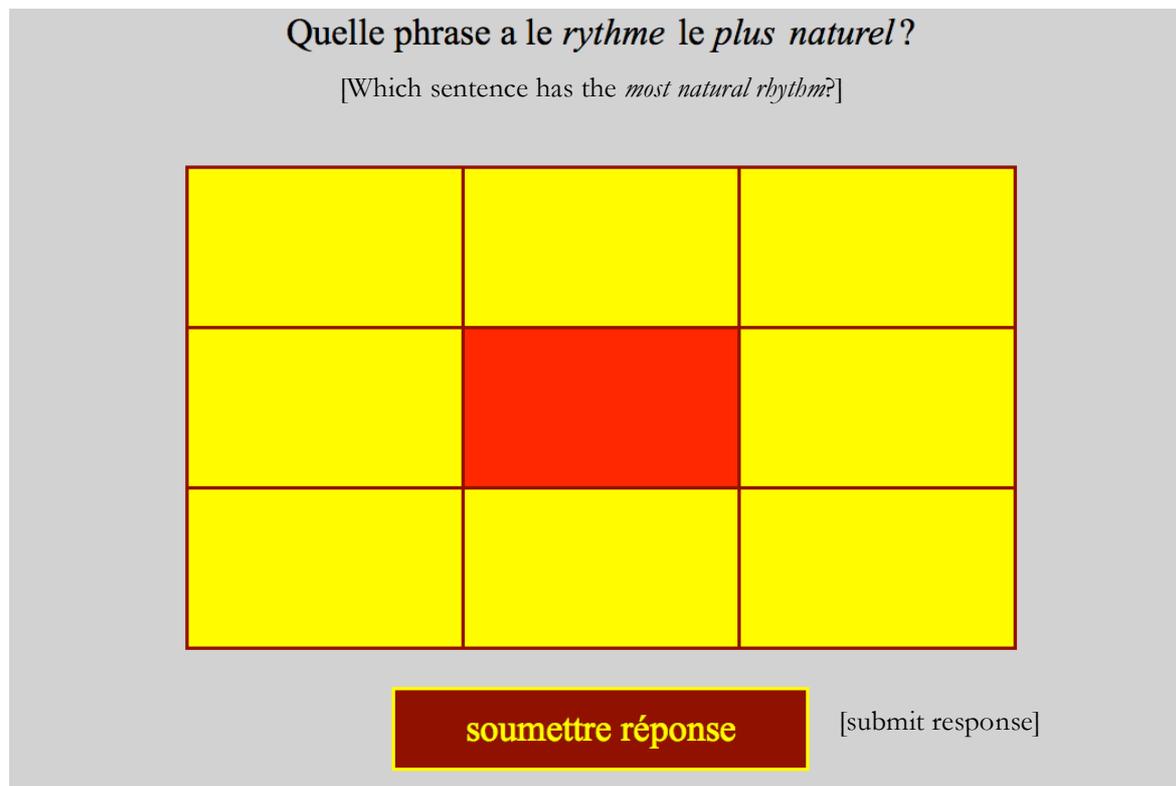
Wie immer isch dDora zschpaat hei cho Comme toujours elle rentrait tard le soir	As always Dora came home too late As always she was coming home late in the evening
Mir wärdet e Rächmig erwarte Nous devons escompter deux quittances	We will expect a bill We have to expect two bills
De Röver mues sMässer verstecke Le voleur doit cacher le couteau	The robber must hide the knife The robber must hide the knife
Im Summer chönd dMäitli go riite En été elles faisaient du cheval	In summer the girls can go horse riding In summer they used to go horse riding
Si würdet gern Tennis go spile Ils voudraient pratiquer le tennis	They would like to play tennis They would like to practise at tennis
De Maaler mues dNaaricht verzele Un artiste devrait dire les nouvelles	The artist must tell the news An artist should tell the news
Im Baanhof isch sRauche verbote Dans la gare on exclut les fumeurs	In the station smoking is forbidden In the station smokers are forbidden
DRoberta hät dMeli verprüglet Raphaëlle a battu Mélanie	Roberta hit Meli Raphaëlle hit Melanie
DFrau Maier gaht dEier go poschte Madame Blanc va payer de bons oeufs	Mrs Maier is going to buy some eggs Mrs Blanc is going to buy some good eggs
De Leerer tuet dSchüeler bestraafe L'enseignante chapitrait l'étudiant	The teacher tells the pupils off The teacher was telling the pupil off
DRebekka gaht hüt a ne Party Rebecca fait la fête aujourd'hui	Rebekka is going to a party today Rebecca is having a party today
DTurische gönd amigs go fische Quelquefois les touristes aiment la pêche	The tourists sometimes go fishing Sometimes the tourists like to go fishing
E Putzfrau tuet dMöbel poliere La cireuse va polir les placards	A cleaning lady polishes the furniture The floor polisher is going to polish the cupboards
Im Lade tüends Öpfel verchaufe L'épicerie vend souvent des pommes mûres	In the shop they sell apples The grocery shop often sells ripe apples
Am Samschtig sind dBuebe go räne Vendredi les garçons ont couru	On Saturday the boys went running On Friday the boys ran

8.3.2 Instructions

Thank you for participating in this study. You will do a listening exercise, and after that I will ask you what you thought about it.

Instructions

Click the mouse anywhere on the screen to start. Then you will see a rectangle of boxes on screen as in this picture:



If you click on each box, you will hear a sentence, and the box will become red. In the rectangle all the sentences have the same words, but they are different in terms of how natural their rhythm sounds. Your task is to choose the sentence which you think has the most natural rhythm. You will be able to listen to each sentence a maximum of three times.

When you have reached your final decision, make sure that the box which you want to choose is red, and then click 'submit response'. After you have clicked to continue, you will see a new rectangle of boxes on the screen. You will start the task again with a newly worded sentence, and you will choose which sentence in the rectangle has the most natural rhythm.

This task will repeat 30 times. You will also have a short practice session (with one sentence) before the main session. You will have chance to take a short break three times when you see the instruction to do so on screen.

Summary

- Choose the sentence which has the most natural rhythm in each rectangle.

If you have any questions, please ask me now.

8.3.3 Post-test questionnaire

Thank you for participating in this study. Now I'd like to know what you thought about it.

In general, how difficult did you find it to judge which sentence had the most natural rhythm?

(scale: 1 = very easy, 7 = very difficult) _____

What does rhythm mean to you? Which criteria did you use to judge a natural rhythm?

Did you have a particular technique for doing the task?

Could you say where the speaker is from? (country/area) _____

Did you notice anything unusual about the speaker, for example, his accent or style?

Please detail your musical training: None

Otherwise _____

Other comments

Thank you for your time 😊

8.3.4 Results

Language			F0 _{Low}	F0 _{Norm}	F0 _{High}
SG ($k=12$)	DUR _{Short}	\bar{x}	0.58	0.83	0.42
		s	0.79	1.19	0.79
	DUR _{Norm}	\bar{x}	3.67	6.33	3.08
		s	2.53	2.74	2.47
	DUR _{Long}	\bar{x}	0.33	0.42	0.33
		s	0.65	0.90	0.65
(S)Fr ($k=12$)	DUR _{Short}	\bar{x}	1.67	2.25	1.33
		s	1.67	1.91	1.92
	DUR _{Norm}	\bar{x}	3.08	4.33	0.92
		s	2.31	3.06	1.08
	DUR _{Long}	\bar{x}	1.08	1.42	0.25
		s	1.08	1.08	0.45

Table 8-7 – Data from Figure 5-9 (preliminary results): mean and standard deviation (across k subjects) of response frequency per stimulus type (out of 18 sentences)

Language			F0 _{Low}	F0 _{Norm}	F0 _{High}
SG ($k=47$)	DUR _{Short}	\bar{x}	1.94	2.02	2.13
		s	1.97	2.07	2.20
	DUR _{Norm}	\bar{x}	6.83	8.81	6.13
		s	2.84	4.11	2.60
	DUR _{Long}	\bar{x}	0.53	0.89	0.72
		s	0.88	1.29	1.08
SFr ($k=50$)	DUR _{Short}	\bar{x}	3.22	4.38	2.18
		s	2.28	2.73	2.40
	DUR _{Norm}	\bar{x}	5.72	6.08	2.64
		s	3.12	3.99	1.88
	DUR _{Long}	\bar{x}	2.70	1.96	1.12
		s	2.57	1.73	1.30
Fr ($k=48$)	DUR _{Short}	\bar{x}	4.50	4.58	2.58
		s	2.39	2.76	2.40
	DUR _{Norm}	\bar{x}	5.27	6.10	1.85
		s	2.89	3.45	1.89
	DUR _{Long}	\bar{x}	2.10	1.98	1.02
		s	1.93	1.83	1.67

Table 8-8 – Data from Figure 5-10 (main experiment results): mean and standard deviation (across k subjects) of response frequency per stimulus type (out of 30 sentences)

Language			F0 _{Low}	F0 _{Norm}	F0 _{High}
northern Fr (<i>k</i> =33)	DUR _{Short}	\bar{x}	4.82	4.85	2.73
		<i>s</i>	2.48	2.61	2.50
	DUR _{Norm}	\bar{x}	5.18	5.64	1.67
		<i>s</i>	3.06	3.43	1.81
	DUR _{Long}	\bar{x}	1.97	1.94	1.21
		<i>s</i>	1.91	1.41	1.92
southern Fr (<i>k</i> =15)	DUR _{Short}	\bar{x}	3.80	4.00	2.27
		<i>s</i>	2.08	3.07	2.19
	DUR _{Norm}	\bar{x}	5.47	7.13	2.27
		<i>s</i>	2.56	3.40	2.05
	DUR _{Long}	\bar{x}	2.40	2.07	0.60
		<i>s</i>	1.99	2.58	0.83

Table 8-9 – Data from Figure 5-12: mean and standard deviation (across *k* subjects) of response frequency per stimulus type (out of 30 sentences)

8.4 Experiment 4 (chapter 6)

8.4.1 Texts

SG: De Biiswind und d Sune

Emaal händ de Biiswind und d Sune gschritte, wèèr vo bäidne das ächt de schtèrcher seig. Da chunt en Maa dethèèr, won en ticke Mantel aaghaa hät. Doo sind s röötig woorde, das dèè de schtèrcher seig, wo dèè Maa dezue bringi, das er sin Mantel abziei. De Biiswind hät aafè blaase so fescht das er hät chöne, aber de Maa hät nu de Mantel änger gnaa. Doo hät d Sune aafè schiine, imer wèèrmer, bis de Maa de Mantel abzoge hät. Doo hät de Biiswind müese zuegèè, das d Sun schtèrcher seig wede èèr. (Fleischer and Schmid 2006)

Fr: La bise et le soleil

La bise et le soleil se disputaient, chacun assurant qu'il était le plus fort. Quand ils ont vu un voyageur qui s'avancait, enveloppé dans son manteau, ils sont tombés d'accord que celui qui arriverait le premier à le lui faire ôter serait regardé comme le plus fort. Alors, la bise s'est mise à souffler de toutes ses forces, mais plus elle soufflait, plus le voyageur serrait son manteau autour de lui. Finalement, elle renonça à le lui faire ôter. Alors, le soleil commença à briller et au bout d'un moment le voyageur, réchauffé, ôta son manteau. Ainsi, la bise dut reconnaître que le soleil était le plus fort. (*Handbook of the IPA* 1999)

8.4.2 Results

Language	Vowel PVIs				Syllable PVIs				
		<i>durational</i>	<i>tonal</i>	<i>combined</i>	<i>weighted</i>	<i>durational</i>	<i>tonal</i>	<i>combined</i>	<i>weighted</i>
SG ($k=10$)	\bar{x}	67.62	81.01	74.04	68.36	58.94	81.01	70.06	61.14
	s	5.03	10.05	5.41	4.46	3.70	10.05	6.02	3.88
SFr ($k=10$)	\bar{x}	49.74	73.14	61.18	62.07	45.17	73.14	58.96	60.02
	s	3.94	2.90	2.81	2.76	3.50	2.90	2.35	2.33
Fr ($k=10$)	\bar{x}	50.13	70.11	60.01	59.32	43.74	70.11	56.82	55.93
	s	3.95	4.88	3.24	3.20	2.64	4.88	3.20	3.12

Table 8-10 – Data from Figure 6-2: mean and standard deviation (across k subjects) of each PVI type

Bibliography

- Abercrombie, D. (1967). *Elements of General Phonetics*. Edinburgh: Edinburgh University Press.
- Adams, C. (1979). *English Speech Rhythm and the Foreign Learner*. The Hague: Mouton.
- Allen, G. D. (1972). Location of rhythmic stress beats. *Language and Speech*, 15, 72-100 and 179-195.
- Allen, G. D. (1973). Segmental timing control in speech production. *Journal of Phonetics*, 1, 219-237.
- Allen, G. D. (1975). Speech rhythm: its relation to performance universals and articulatory timing. *Journal of Phonetics*, 3, 75-86.
- Allen, G. D., & Hawkins, S. (1979). Trochaic rhythm in children's speech. In H. Hollien & P. Hollien (Eds.), *Current Issues in the Phonetic Sciences* (pp. 927-933). Amsterdam: John Benjamins.
- Allen, G. D., & Hawkins, S. (1980). Phonological Rhythm: Definition and Development. In G. H. Yeni-Komshian, J. F. Kavanagh & C. A. Ferguson (Eds.), *Child Phonology: Production* (Vol. 1, pp. 227-256). New York: Academic Press.
- Ambühl, D., von Kaenel, V., & Peter, C. (Eds.). (2003). Proceedings of *Sprachen und Kulturen: viersprachig, mehrsprachig, vielsprachig / Langues et cultures: la Suisse, un pays où l'on parle quatre langues...et plus. Biel-Bienne, Switzerland, 14th November 2002*. Bern: Schweizerische Akademie der Geistes- und Sozialwissenschaften.
- Andreassen, H. N. (2006). Aspects de la durée vocalique dans le vaudois. *Bulletin PFC: Phonologie du Français Contemporain – Usages, Variétés et Structure*, 6, 115-134.
- Andreassen, H. N., & Detey, S. (2007). Conversation à Nyon: description illustrée d'une variété spécifique de français parlé suisse. *Bulletin PFC: Phonologie du Français Contemporain – Usages, Variétés et Structure*, 7, 109-120.
- Arès, G. (1994). *Parler suisse, parler français*. Vevey: Editions de l'Aire.
- Armstrong, L. E., & Ward, I. C. (1926). *A handbook of English intonation*. Leipzig: Teubner.
- Artésano, C., Di Cristo, A., & Hirst, D. (1995). *Discourse-based empirical evidence for a multi-class accent system in French*. Paper presented at the 13th International Congress of Phonetic Sciences, Stockholm, Sweden, 13th-19th August 1995.
- Arvaniti, A. (1994). Acoustic features of Greek rhythmic structure. *Journal of Phonetics*, 22, 239-268.
- Arvaniti, A. (2009). Rhythm, timing, and the timing of rhythm. *Phonetica*, 66, 46-63.

- Auer, P. (1993). *Is a rhythm-based typology possible? A study of the role of prosody in phonological typology*. Unpublished doctoral thesis, University of Freiburg, Freiburg.
- Auer, P., & Couper-Kuhlen, E. (1994). Rhythmus und Tempo konversationeller Alltagssprache. *Zeitschrift für Literaturwissenschaft und Linguistik*, 96, 78-106.
- Auer, P., Couper-Kuhlen, E., & Müller, F. (1999). *Language in Time*. New York and Oxford: Oxford University Press.
- Auer, P., Gilles, P., Peters, J., & Selting, M. (2000). Intonation regionaler Varietäten des Deutschen: Vorstellung eines Forschungsprojekts. In D. Stellmacher (Ed.), *Dialektologie zwischen Tradition und Neuansätzen. Beiträge der internationalen Dialektologentagung, Göttingen, 19th-21st October 1998* (pp. 222-239). Stuttgart: Steiner.
- Baayen, R. H. (2008). *Analyzing Linguistic Data: A practical introduction to statistics*. Cambridge: Cambridge University Press.
- Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4), 390-412.
- Bailey, T. M., Plunkett, K., & Scarpa, E. (1999). A cross-linguistic study in learning prosodic rhythms: rules, constraints, and similarity. *Language and Speech*, 42(1), 1-38.
- Ball, R. (1997). *The French-speaking world: a practical introduction to sociolinguistic issues*. London: Routledge.
- Barbosa, P. A. (2002). *Explaining cross-linguistic rhythmic variability via a coupled-oscillator model of rhythm production*. Paper presented at Speech Prosody 2002, Aix-en-Provence, France, 11th-13th April 2002.
- Barbosa, P. A. (2007). From syntax to acoustic duration: A dynamical model of speech rhythm production. *Speech Communication*, 49(9), 725-742.
- Barker, G. (2005). *Intonation patterns in Tyrolean German: an Autosegmental-metrical analysis*. New York: Peter Lang.
- Barry, W. J. (2007). Rhythm as an L2 problem: How prosodic is it? In J. Trouvain & U. Gut (Eds.), *Non-native Prosody: phonetic description and teaching practice* (pp. 97-120). Berlin: Mouton de Gruyter.
- Barry, W. J., Andreeva, B., & Koreman, J. (2009). Do rhythm measures reflect perceived rhythm? *Phonetica*, 66, 78-94.
- Barry, W. J., Andreeva, B., Russo, M., Dimitrova, S., & Kostadinova, T. (2003). *Do rhythm measures tell us anything about language type?* Paper presented at the 15th International Congress of Phonetic Sciences, Barcelona, Spain, 3rd-9th August 2003.

- Barry, W. J., van Dommelen, W., & Koreman, J. (2005). Phonetic knowledge in speech technology – and phonetic knowledge from speech technology? In W. J. Barry & W. van Dommelen (Eds.), *The Integration of Phonetic Knowledge in Speech Technology* (pp. 1-12). Dordrecht: Springer.
- Batliner, A., & Möbius, B. (2005). Prosodic models, automatic speech understanding, and speech synthesis: Towards the common ground? In W. J. Barry & W. van Dommelen (Eds.), *The Integration of Phonetic Knowledge in Speech Technology* (pp. 21-44). Dordrecht: Springer.
- Bayard, C., & Jolivet, R. (1984). Des Vaudois devant la norme. *Le français moderne*, 52, 151-158.
- Beaugendre, F. (1994). *Une étude perceptive de l'intonation du français*. Unpublished doctoral thesis, University of Paris XI, Orsay, France.
- Beckman, M. E. (1992). Evidence for speech rhythms across languages. In Y. Tōkura, E. Vatikiotis-Bateson & Y. Sagisaka (Eds.), *Speech Perception, Production and Linguistic Structure* (pp. 457-463). Tokyo: OHM Publishing Co.
- Beckman, M. E. (1993). Modeling the production of prosody. *Department of Linguistics and Phonetics, Lund University, Working papers*, 41, 258-261.
- Beckman, M. E., & Pierrehumbert, J. (1986). Intonational structure in Japanese and English. *Phonology Yearbook*, 3, 255-309.
- Beddor, P. S., & Gottfried, T. L. (1995). Methodological issues in cross-language speech perception research with adults. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp. 207-232). Baltimore: York Press.
- Beilstein-Schaufelberger, A. (2007). *Züritüütsch: Schweizerdeutsch* (2nd ed.). Zürich: Karl Schwegler AG, Grafischer Betrieb & Verlag.
- Benguerel, A.-P. (1971). Duration of French vowels in unemphatic stress. *Language and Speech*, 14, 383-391.
- Benguerel, A.-P. (1973). Corrélats physiologiques de l'accent en français. *Phonetica*, 27, 21-35.
- Benguerel, A.-P. (1986). On the measurement of rhythmic irregularity. *Journal of Phonetics*, 14, 325-327.
- Benguerel, A.-P., & D'Arcy, J. (1986). Time-warping and the perception of rhythm in speech. *Journal of Phonetics*, 14, 231-246.
- Bertinetto, P. M. (1980). The perception of stress by Italian speakers. *Journal of Phonetics*, 8, 385-395.
- Bertinetto, P. M., & Bertini, C. (2008). *On modeling the rhythm of natural languages*. Paper presented at Speech Prosody 2008, Campinas, Brazil, 6th-9th May 2008.

- Bertinetto, P. M., & Fowler, C. A. (1989). On sensitivity to durational modifications in Italian and English. *Rivista di Linguistica*, 1, 69-94.
- Bertoncini, J., Floccia, C., Nazzi, T., & Mehler, J. (1995). Morae and syllables: Rhythmical basis of speech representations in neonates. *Language and Speech*, 38(4), 311-329.
- Best, C. T. (1995). A direct realist perspective on cross-language speech perception. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp. 167-200). Baltimore: York Press.
- Blevins, J. (1995). The syllable in phonological theory. In J. A. Goldsmith (Ed.), *The Handbook of Phonological Theory* (pp. 206-244). Oxford: Blackwell.
- Bloch, B. (1950). Studies in colloquial Japanese IV: phonemics. *Language*, 26, 86-125.
- Boersma, P. (2005). *PitchTier: Stylize...* Page from the Praat manual <http://www.fon.hum.uva.nl/praat/manual/PitchTier__Stylize____.html> [accessed 26th November 2008]
- Boersma, P. (2008). *What happens during duration manipulation?* Question answered in the Praat online user group <<http://uk.groups.yahoo.com/group/praat-users/message/3526>> [accessed 1st July 2008]
- Boersma, P., & Weenik, D. (2008-2009). *Praat: doing phonetics by computer* (versions 5.0.05, 5.0.21, 5.1.01, 5.1.04) [Computer program] <<http://www.praat.org/>> [accessed 20th January 2008 to 10th April 2009 – updated versions]
- Bolinger, D. (1965). Pitch accent and sentence rhythm. In I. Abe & T. Kanekiyo (Eds.), *Forms of English: accent, morpheme, order* (pp. 139-180). Harvard: Harvard University Press.
- Bolton, T. L. (1894). Rhythm. *The American Journal of Psychology*, 6(2), 145-238.
- Boucher, V. J. (2006). On the function of stress rhythms in speech: Evidence of a link with grouping effects on serial memory. *Language and Speech*, 49(4), 495-519.
- Boudreault, M. (1968). *Rythme et mélodie de la phrase parlée en France et au Québec*. Québec: Les Presses de l'Université Laval, Paris: Klincksieck.
- Boula de Mareüil, P., & Vieru-Dimulescu, B. (2006). The contribution of prosody to the perception of foreign accent, *Phonetica*, 63, 247-267.
- Bouwhuis, D. G., Bergmans, J., & de Rooij, J. J. (1978). A model for the perception of prosodic boundaries. *IPO Annual Progress Report*, 13, 76-82.
- Bremer, O. (1893). *Deutsche Phonetik*. Leipzig: Breitkopf and Härtel.
- Brown, W. (1908). *Time in English Verse Rhythm*. New York: The Science Press.

- Brown, W. (1911). Studies from the psychological laboratory of the University of California. Temporal and accentual rhythm. *Psychological Review*, 18(5), 336-346.
- Brücke, E. (1871). *Physiologische Grundlagen der neuhochdeutschen Verkunst*. Vienna: Gerold.
- Brunellière, A., Dufour, S., Nguyen, N., & Frauenfelder, U. H. (2009). Behavioral and electrophysiological evidence for the impact of regional variation on phoneme perception. *Cognition*, 111(3), 390-396.
- Carlson, R., & Swerts, M. (2003). Relating perceptual judgments of upcoming prosodic breaks to F0 features. *PHONUM*, 9, 181-184.
- Carlson, R., Granström, B., Heldner, M., House, D., Megyesi, B., Strangert, E., et al. (2002). Boundaries and groupings – the structuring of speech in different communicative situations: a description of the GROG project. Proceedings of Fonetik 2002, *TMH-QPSR (KTH Department of Speech, Music and Hearing: Quarterly Progress and Status Report)*, 43.
- Carter, P. M. (2005). Quantifying rhythmic differences between Spanish, English, and Hispanic English. In R. S. Gess & E. J. Rubin (Eds.), *Theoretical and Experimental Approaches to Romance Linguistics: Selected Papers from the 34th Linguistic Symposium on Romance Languages* (pp. 63-75). Amsterdam and Philadelphia: John Benjamins.
- Carton, F. (1987). Les accents régionaux. In G. Vermès & J. Boutet (Eds.), *France, Pays Multilingue. Tome 2: Pratique des Langues en France* (pp. 29-49). Paris: L'Harmattan.
- Carton, F., Rossi, M., Autesserre, D., & Léon, P. (1983). *Les accents des Français*. Paris: Hachette.
- Catford, J. C. (1988). *A Practical Introduction to Phonetics*. Oxford: Oxford University Press.
- Christen, H. (1998). Convergence and divergence in the Swiss German dialects. *Folia Linguistica*, 32(1-2), 53-68.
- Classe, A. (1939). *The Rhythm of English Prose*. Oxford: Blackwell.
- Classen, K., Dogil, G., Jessen, M., Marasek, K., & Wokurek, W. (1998). Stimmqualität als Korrelat der Wortbetonung im Deutschen. *Linguistische Berichte*, 174, 202-245.
- Collier, R. (1993). On the communicative function of prosody: some experiments. *IPO Annual Progress Report*, 28, 67-75.
- Cooper, A. M., Whalen, D. H., & Fowler, C. A. (1986). P-centers are unaffected by phonetic categorization. *Perception and Psychophysics*, 39(3), 187-196.
- Cooper, W. E., & Sorenson, J. M. (1977). Fundamental frequency contours at syntactic boundaries. *Journal of the Acoustical Society of America*, 62, 682-692.
- Corriveau, K. H., & Goswami, U. (2009). Rhythmic motor entrainment in children with speech and language impairments: Tapping to the beat. *Cortex*, 45(1), 119-130.

- Corriveau, K. H., Pasquini, E. S., & Goswami, U. (2007). Basic auditory processing skills and Specific Language Impairment: A new look at an old hypothesis. *Journal of Speech, Language, and Hearing Research, 50*, 647-666.
- Couper-Kuhlen, E. (1993). *English Speech Rhythm: Form and Function in Everyday Verbal Interaction*. Amsterdam and Philadelphia: John Benjamins.
- Coustenoble, H., & Armstrong, L. (1934). *Studies in French Intonation*. Cambridge: W. Heffer and Sons.
- Crompton, A. (1980). Timing patterns in French. *Phonetica, 37*, 205-234.
- Crowder, R. G. (1982). The demise of short-term memory. *Acta Psychologica, 50*, 75-84.
- Cruttenden, A. (1986). *Intonation*. Cambridge: Cambridge University Press.
- Crystal, D. (1969). *Prosodic Systems and Intonation in English*. Cambridge: Cambridge University Press.
- Crystal, D. (1970). Prosodic systems and language acquisition. In P. R. Léon, G. Faure & A. Rigault (Eds.), *Prosodic Feature Analysis* (pp. 77-90). Montreal: Didier.
- Crystal, D. (1973). Linguistic mythology and the first year of life. *British Journal of Disorders of Communication, 8*(1), 29-36.
- Crystal, D. (Ed.) (1985) *Dictionary of Linguistics and Phonetics* (2nd ed.). Oxford: Blackwell.
- Cummins, F. (2002). *Speech Rhythm and Rhythmic Taxonomy*. Paper presented at Speech Prosody 2002, Aix-en-Provence, France, 11th-13th April 2002.
- Cummins, F., & Port, R. F. (1998). Rhythmic constraints on stress timing in English. *Journal of Phonetics, 26*, 145-171.
- Cutler, A. (2005). Lexical Stress. In D. B. Pisoni & R. E. Remez (Eds.), *Handbook of Speech Perception* (pp. 264-289). Oxford: Blackwell.
- Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language, 31*, 218-236.
- Cutler, A., & Norris, D. G. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance, 14*, 113-121.
- Cutler, A., Dahan, D., & van Donselaar, W. (1997). Prosody in the comprehension of spoken language: A literature review. *Language and Speech, 40*(2), 141-202.
- Cutler, A., Mehler, J., Norris, D., & Segui, J. (1986). The syllable's differing role in the segmentation of French and English. *Journal of Memory and Language, 25*, 385-400.

- Cutler, A., Mehler, J., Norris, D., & Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology*, 24, 381-410.
- Dahan, D. (1996). *The role of rhythmic groups in the segmentation of continuous French speech*. Paper presented at the 4th International Conference on Spoken Language, Philadelphia, USA, 3rd-6th October 1996.
- Dasher, R., & Bolinger, D. (1982). On pre-accentual lengthening. *Journal of the International Phonetic Association*, 12, 58-69.
- Dauer, R. (1983). Stress-timing and syllable-timing reanalyzed. *Journal of Phonetics*, 11, 51-62.
- Dauer, R. (1987). *Phonetic and Phonological Components of Language Rhythm*. Paper presented at the 11th International Congress of Phonetic Sciences, Tallinn, Estonia, 1st-7th August 1987.
- de Bot, K. (1986). The transfer of intonation and the missing data base. In E. Kellerman & M. Sharwood-Smith (Eds.), *Cross-linguistic Influence in Second Language Acquisition* (pp. 110-119). New York and Oxford: Pergamon Institute of English.
- de Pijper, J. R., & Sanderman, A. A. (1994). On the perceptual strength of prosodic boundaries and its relation to suprasegmental cues. *Journal of the Acoustical Society of America*, 96(4), 2037-2047.
- de Rooij, J. J. (1976). Perception of prosodic boundaries. *IPO Annual Progress Report*, 11, 20-24.
- Delais-Roussarie, E., Riolland, A., Doetjes, J., & Marandin, J. M. (2002). *The Prosody of Post-focus Sequences in French*. Paper presented at Speech Prosody 2002, Aix-en-Provence, France, 11th-13th April 2002.
- Delattre, P. (1966a). Les dix intonations de base du français. *The French Review*, 40(1), 1-14.
- Delattre, P. (1966b). *Studies in French and Comparative Phonetics: English, German, Spanish, and French*. Heidelberg: Groos.
- Dell, F. (1984). L'accentuation dans les phrases en français. In F. Dell, D. Hirst & J.-R. Vergnaud (Eds.), *Forme sonore du langage: Structure des représentations en phonologie* (pp. 65-122). Paris: Hermann.
- Dellwo, V. (2006). Rhythm and speech rate: A variation coefficient for deltaC. In P. Karnowski & I. Sziget (Eds.), *Language and Language Processing: Proceedings of the 38th Linguistic Colloquium* (pp. 231-241). Frankfurt: Peter Lang.
- Derwing, T. M., & Munro, M. J. (2009). Putting accent in its place: Rethinking obstacles to communication. *Language Teaching*, 42(4), 476-490.
- Derwing, T. M., Munro, M. J., & Wiebe, G. E. (1998). Evidence in favor of a broad framework for pronunciation instruction. *Language Learning*, 48, 393-410.

- Deterding, D. (1994). *The rhythm of Singapore English*. Paper presented at the 5th Australian International Conference on Speech Science and Technology, Western Australia, 6th-8th December 1994.
- Deterding, D. (2001). The measurement of rhythm: a comparison of Singapore and British English. *Journal of Phonetics*, 29, 217-230.
- Di Cristo, A. (1998). Intonation in French. In D. Hirst & A. Di Cristo (Eds.), *Intonation Systems: A Survey of Twenty Languages* (pp. 195-218).
- Di Cristo, A. (1999). Vers une modélisation de l'accentuation du français: première partie. *French Language Studies*, 9, 143-179.
- Di Cristo, A. (2000). Vers une modélisation de l'accentuation du français: deuxième partie. *French Language Studies*, 10, 27-44.
- Di Cristo, A., & Hirst, D. (1993). Rythme syllabique, rythme mélodique et représentation hiérarchique de la prosodie du français. *Travaux de l'Institut de Phonétique d'Aix*, 15, 9-24.
- Di Cristo, A., & Hirst, D. J. (1986). Modelling French micromelody: Analysis and synthesis. *Phonetica*, 43, 11-30.
- Di Cristo, A., & Hirst, D. J. (1997). L'accentuation non-emphatique en français: stratégies et paramètres. In J. Perrot (Ed.), *Polyphonie pour Iván Fónagy: mélanges offerts en hommage à Iván Fónagy* (pp. 71-101). Paris: L'Harmattan.
- Dictionnaire Suisse Romande: particularités lexicales du français contemporain; une contribution au Trésor des vocabulaires francophones*. (2004). Ed. by A. Thibault with P. Knecht, G. Boeri & S. Quenet. Geneva: Zoé.
- Dogil, G. (1999). Slavic Languages. In H. van der Hulst (Ed.), *Word Prosodic Systems in the Languages of Europe*. Berlin and New York: Mouton de Gruyter.
- Donegan, P., & Stampe, D. (1983). Rhythm and the holistic organization of language structure. In F. Richardson (Ed.), *Papers from the CLS Parasession on the Interplay of Phonology, Morphology & Syntax* (pp. 337-353). Chicago: CLS.
- Donovan, A., & Darwin, C. J. (1979). *The Perceived Rhythm of Speech*. Paper presented at the 9th International Congress of Phonetic Sciences, Copenhagen, Denmark, 6th-11th August 1979.
- Duarte, D., Galves, A., Garcia, N. L., & Maronna, R. (2001). *The statistical analysis of acoustic correlates of speech rhythm*. Paper presented at the Workshop on Rhythmic Patterns, Parameter Setting and Language Change, ZiF, University of Bielefeld.
<<http://www.physik.uni-bielefeld.de/complexity/duarte.pdf> 2001>

- Duckworth, J. (1967). The rhythm of spoken English. In D. C. Wigglesworth (Ed.), *Selected Conference Papers of the Association of Teachers of English as a Second Language*. Los Altos, California: Language Research Associates' Press.
- Dufour, S., Nguyen, N., & Frauenfelder, U. H. (2007). The perception of phonemic contrasts in a non-native dialect. *Journal of the Acoustical Society of America*, *121*, EL131-EL136.
- Dupoux, E., Pallier, C., Sebastián-Gallés, N., & Mehler, J. (1997). A destressing 'deafness' in French? *Journal of Memory and Language*, *36*, 406-421.
- Dupoux, E., Peperkamp, S., & Sebastián-Gallés, N. (2001). A robust method to study stress 'deafness'. *Journal of the Acoustical Society of America*, *110*(3), 1608-1618.
- Dupoux, E., Sebastián-Gallés, N., Navarrete, E., & Peperkamp, S. (2008). Persistent stress 'deafness': The case of French learners of Spanish. *Cognition*, *106*(2), 682-706.
- Dupoux, E., Peperkamp, S., & Sebastián-Gallés, N. (2010). Limits on bilingualism revisited: Stress 'deafness' in simultaneous French-Spanish bilinguals. *Cognition*, *114*(2), 266-275.
- Durand, J., Laks, B., & Lyche, C. (2003). Le projet 'Phonologie du Français Contemporain' (PFC). *La Tribune Internationale des Langues Vivantes*, *33*, 3-9.
- Echols, C. H., Crowhurst, M. J., & Childers, J. B. (1997). The perception of rhythmic units in speech by infants and adults. *Journal of Memory and Language*, *36*, 202-225.
- Eimas, P. D. (1963). The relation between identification and discrimination along speech and non-speech continua. *Language and Speech*, *6*, 206-217.
- Engstrand, O. (1987). Durational patterns of Lule Sami phonology. *Phonetica*, *44*, 117-128.
- Engstrand, O., & Krull, D. (2002). Duration of syllable-sized units in casual and elaborated speech: cross-language observations on Swedish and Spanish. *TMH-QPSR (KTH Department of Speech, Music and Hearing: Quarterly Progress and Status Report)*, *44*, 69-72.
- Faraway, J. J. (2006). *Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models*. London: Chapman and Hall/CRC Press.
- Faure, G., Hirst, D. J., & Chafcouloff, M. (1980). Rhythm in English: Isochronism, pitch and perceived stress. In L. R. Waugh & C. H. van Schooneveld (Eds.), *The Melody of Language* (pp. 71-79). Baltimore: University Park Press.
- Ferragne, E. (2008). *Étude phonétique des dialectes modernes de l'anglais des Îles Britanniques: vers l'identification automatique du dialecte*. Unpublished doctoral thesis, Université Lyon 2, Lyon.
- Field, A. (2005). *Discovering Statistics Using SPSS*. London: SAGE.
- Fitch, W. T., Hauser, M. D., & Chomsky, N. (2005). The evolution of the language faculty: Clarifications and implications. *Cognition*, *97*, 179-210.

- Fitzpatrick-Cole, J. (1999). *The Alpine Intonation of Bern Swiss German*. Paper presented at the 14th International Congress of Phonetic Sciences, San Francisco, USA, 1st-7th August 1999.
- Flege, J. E. (1993). Production and perception of a novel, second-language phonetic contrast. *Journal of the Acoustical Society of America*, 93(3), 1589-1607.
- Flege, J. E. (1995). Second-language speech learning: Theory, findings, and problems. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research*. Baltimore: York Press.
- Flege, J. E., MacKay, I. R. A., & Meador, D. (1999). Native Italian speakers' perception and production of English vowels. *Journal of the Acoustical Society of America*, 106(5), 2973-2987.
- Fleischer, J., & Schmid, S. (2006). Illustrations of the IPA: Zurich German. *Journal of the International Phonetic Association*, 36(2), 243-253.
- Fletcher, J. (1991). Rhythm and final lengthening in French. *Journal of Phonetics*, 19, 193-212.
- Fletcher, J. (2010). The prosody of speech: Timing and rhythm. In W. J. Hardcastle, J. Laver & F. E. Gibbon (Eds.), *Handbook of the Phonetic Sciences* (2nd ed.) (pp. 523-602). Blackwell: Oxford.
- Fónagy, I. (1949). *Displacement of the 'accent d'intensité' in the Romance languages (written in Hungarian)*. Unpublished doctoral thesis, Universite L. Eötvös, Budapest.
- Fónagy, I. (1980). L'accent français: accent probilitaire. In P. R. Léon & M. Rossi (Eds.), *L'accent en français contemporain: Studia Phonetica*, 15 (pp. 123-233). Montréal: Didier.
- Fouché, P. (1933-34). L'évolution phonétique du français du XVI^{ème} siècle à nos jours. *Le français moderne*, 1-2, 217-236.
- Fougeron, C., & Jun, S.-A. (1998). Rate effects on French intonation: prosodic organization and phonetic realization. *Journal of Phonetics*, 26, 45-69.
- Fowler, C. A. (1977). *Timing Control in Speech Production*. Bloomington: Indiana University Linguistics Club.
- Fowler, C. A. (1979). "Perceptual centres" in speech production and perception. *Perception and Psychophysics*, 25, 375-388.
- Fox, A. (2000). *Prosodic Features and Prosodic Structure: The Phonology of Suprasegmentals*. Oxford: Oxford University Press.
- Fox, R. A. (1985). Auditory contrast and speaker quality variation in vowel perception. *Journal of the Acoustical Society of America*, 77(4), 1552-1559.
- Fox, R. A., & Lehiste, I. (1987). The effect of vowel quality variations on stress-beat location. *Journal of Phonetics*, 15, 1-13.

- Fraisse, P. (1982). Rhythm and Tempo. In D. Deutsch (Ed.), *Psychology of Music* (pp. 149-180). New York and London: Academic Press.
- Fraisse, P. (1984). Perception and estimation of time. *Annual Review of Psychology*, 35, 1-36.
- Frankish, C. (1985). Modality-specific grouping effects in short-term memory. *Journal of Memory and Language*, 24, 200-209.
- Frankish, C. (1989). Perceptual organization and pre-categorical acoustic storage. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 15(3), 469-479.
- Frey, E. (1975). *Stuttgarter Schwäbisch*. Marburg: Elwert.
- Frota, S., Vigário, M., & Martins, F. (2002). *Language Discrimination and Rhythm Classes: Evidence from Portuguese*. Paper presented at Speech Prosody 2002, Aix-en-Provence, France, 11th-13th April 2002.
- Fry, D. B. (1955). Duration and intensity as physical correlates of linguistic stress. *Journal of the Acoustical Society of America*, 27, 765-768.
- Fry, D. B. (1958). Experiments in the perception of stress. *Language and Speech*, 1, 126-152.
- Fry, D. B. (1965). *The dependence of stress judgments on vowel formant structure*. Paper presented at the 5th International Congress of Phonetic Sciences, Münster, Germany 1965.
- Fudge, E. C. (1969). Syllables. *Journal of Linguistics*, 5(2), 253-286.
- Fujisaki, H., & Hirose, K. (1984). Analysis of voice fundamental frequency contours for declarative sentences of Japanese. *Journal of the Acoustical Society of Japan* 5(4), 233-241.
- Gadet, F. (2007). *La variation sociale en français*. Paris: Ophrys.
- Galloway, R. E. (2007). *Bilinguals' interacting phonologies? A study of speech production in French~Swiss German bilinguals*. Unpublished MPhil thesis, University of Cambridge, Cambridge.
- Galves, A., Garcia, J., Duarte, D., & Galves, C. (2002). *Sonority as a basis for rhythmic class discrimination*. Paper presented at Speech Prosody 2002, Aix-en-Provence, France, 11th-13th April 2002.
- Garson, G. D. (2009). *Logistic regression*. Statnotes: Topics in Multivariate Analysis. <<http://faculty.chass.ncsu.edu/garson/pa765/statnote.htm>> [accessed April 2009]
- Ghitza, O., & Greenberg, S. (2009). On the possible role of brain rhythms in speech perception: Intelligibility of time-compressed speech with periodic and aperiodic insertions of silence. *Phonetica*, 66, 113-126.
- Gibbon, D. (1998). Intonation in German. In D. Hirst & A. Di Cristo (Eds.), *Intonation Systems: A Survey of Twenty Languages* (pp. 78-95). Cambridge: Cambridge University Press.

- Gibbon, D., & Gut, U. (2001). *Measuring Speech Rhythm*. Paper presented at Eurospeech 2001, Aalborg, Denmark, 3rd-7th September 2001.
- Gill, A. (1936). Remarques sur l'accent tonique en français contemporain. *Le Français Moderne*, 4, 311-318.
- Goswami, U., Gerson, D., & Astruc, L. (2009). Amplitude envelope perception, phonology and prosodic sensitivity in children with developmental dyslexia. *Reading and Writing: online journal* <<http://www.springerlink.com/content/eqk2318618670m36/fulltext.pdf>>.
- Grabe, E. (2002). *Variation Adds to Prosodic Typology*. Paper presented at Speech Prosody 2002, Aix-en-Provence, France, 11th-13th April 2002.
- Grabe, E., & Low, E. L. (2002). Durational variability in speech and the Rhythm Class Hypothesis. In N. Warner & C. Gussenhoven (Eds.), *Papers in Laboratory Phonology VII* (pp. 515-543). Berlin: Mouton de Gruyter.
- Grabe, E., Gut, U., Post, B., & Watson, I. M. C. (1999). *The Acquisition of Rhythm in English, French and German*. Paper presented at The Child Language Seminar, London, UK 1999.
- Grabe, E., Post, B., & Watson, I. M. C. (1999). *The Acquisition of Rhythmic Patterns in English and French*. Paper presented at the 14th International Congress of Phonetic Sciences, San Francisco, USA, 1st-7th August 1999.
- Grammont, M. (1914). *Traité pratique de prononciation française* (1st ed.). Paris: Delagrave.
- Grosjean, F., Carrard, S., Godio, C., & Grosjean, L. (2007). Long and short vowels in Swiss French: their production and perception. *French Language Studies*, 17, 1-19.
- Grover, C. N. & Terken, J. M. B. (1995). The role of stress and accent in the perception of speech rhythm. *IPO Annual Progress Report*, 30, 30-37.
- Guaitella, I. (1999). Rhythm in speech: What rhythmic organizations reveal about cognitive processes in spontaneous speech production versus reading aloud. *Journal of Pragmatics*, 31, 509-523.
- Gussenhoven, C. (1984). *On the Grammar and Semantics of Sentence Accents*. Dordrecht: Foris.
- Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. Cambridge: Cambridge University Press.
- Gussenhoven, C., & Blom, J. G. (1978). Perception of prominence by Dutch listeners. *Phonetica*, 35, 216-230.
- Gutknecht, C. (1972). *A note on the role of pitch as an element of the accent within synthetic pairs of syllables*. Paper presented at the 7th International Congress of Phonetic Sciences, Montreal, Quebec, 22nd-28th August 1971.

- Haas, W. (1982). Die deutschsprachige Schweiz. In R. Schläpfer (Ed.), *Die vier-sprachige Schweiz* (pp. 73-160). Zürich and Cologne: Benziger.
- Hall, A. (1946). Colloquial French phonology. *Studies in Linguistics*, 4, 70-79.
- Halle, M., & Vergnaud, J.-R. (1987). *An Essay on Stress*. Cambridge, MA: MIT Press.
- Halliday, M. A. K. (1960). Categories of the theory of grammar. *Word*, 17, 241-292.
- Halliday, M. A. K. (1970). *A Course in Spoken English: Intonation*. Oxford: Oxford University Press.
- Hämäläinen, J. A., Leppänen, P. H. T., Eklund, K., Thomson, J., Richardson, U., Guttorm, T. K., et al. (2009). Common variance in amplitude envelope perception tasks and their impact on phoneme duration perception and reading and spelling in Finnish children with reading disabilities. *Applied Psycholinguistics*, 30, 511-530.
- Handbook of the International Phonetic Association*. (1999). Cambridge: Cambridge University Press.
- Harsin, C. A. (1997). Perceptual-center modeling is affected by including acoustic rate-of-change modulations. *Perception and Psychophysics*, 59(2), 243-251.
- Häsler, K., Hove, I., & Siebenhaar, B. (2005). Die Prosodie des Schweizerdeutschen – Erkenntnisse aus der sprachsynthetischen Modellierung von Dialekten. *Linguistik Online*, 24(3), 187-224.
- Haug, W. (2002). *Neue Dynamik in der Sprachenlandschaft Schweiz*. Paper presented at 'Sprachen und Kulturen: viersprachig, mehrsprachig, vielsprachig. Langues et cultures: la Suisse, un pays où l'on parle quatre langues....et plus', Biel-Bienne, Switzerland, 14th November 2002.
- Haugen, E. (1966). Dialect, language, nation. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 97-116). Harmondsworth: Penguin.
- Hawkins, R. (1993). Regional variation in France. In C. Sanders (Ed.), *French Today* (pp. 55-84). Cambridge: Cambridge University Press.
- Hawkins, S. (1999). Auditory capacities and phonological development: Animal, baby, and foreign listeners (chapter 13); Looking for invariant correlates of linguistic units: Two classical theories of speech perception (chapter 14); Re-evaluating assumptions about speech perception: Interactive and integrative theories (chapter 15). In J. M. Pickett (Ed.), *The Acoustics of Speech Communication: Fundamentals, Speech Perception Theory, and Technology* (pp. 183-288). Needham Heights, MA: Allyn and Bacon.
- Hawkins, S., & Smith, R. (2001). Polysp: a polysystemic, phonetically-rich approach to speech understanding. *Rivista di Linguistica*, 13, 99-188.

- Hay, J. F., & Diehl, R. L. (1999). *Effect of duration, intensity and F0 alternations on rhythmic grouping*. Paper presented at the 14th International Congress of Phonetic Sciences San Francisco, USA, 1st-7th August 1999.
- Hay, J. F., & Diehl, R. L. (2007). Perception of rhythmic grouping: Testing the iambic/trochaic law. *Perception and Psychophysics*, 69, 113-122.
- Hayes, B. (1981). *A Metrical Theory of Stress Rules*. Unpublished doctoral thesis, MIT.
- Hirschfeld, U., & Trouvain, J. (2007). Teaching prosody in German as a foreign language. In J. Trouvain & U. Gut (Eds.), *Non-native Prosody: Phonetic Description and Teaching Practice* (pp. 171-188). Berlin: Mouton de Gruyter.
- Hirschfeld, U., & Ulbrich, C. (2002). Untersuchungen zu prosodischen Merkmalen der Standardaussprachen der Bundesrepublik Deutschland und der deutschsprachigen Schweiz. In W. J. Barry & M. Pützer (Eds.), *Festschrift für Max Mangold zum 80. Geburtstag* (pp. 103-128). Saarbrücken.
- Hirst, D. (1998). Intonation in British English. In D. Hirst & A. Di Cristo (Eds.), *Intonation Systems: A Survey of Twenty Languages* (pp. 56-77). Cambridge: Cambridge University Press.
- Hirst, D., & Di Cristo, A. (1984). French intonation: A parametric approach. *Die Neuen Sprachen*, 83(5), 554-569.
- Hoequist, C. (1983a). Durational correlates of linguistic rhythm categories. *Phonetica*, 40, 19-31.
- Hoequist, C. (1983b). Syllable duration in stress-, syllable- and mora-timed Languages. *Phonetica*, 40, 203-237.
- Hoequist, C. (1983c). The Perceptual Center and rhythm categories. *Language and Speech*, 26(4), 367-376.
- Holmes, J., & Holmes, W. (2001). *Speech Synthesis and Recognition* (2nd ed.). New York: Taylor & Francis.
- Hornik, K. (2008). *Frequently Asked Questions on R* <http://cran.r-project.org/doc/FAQ/R-FAQ.html#Why-are-p_002dvalues-not-displayed-when-using-lmer_0028_0029_003f> [accessed 19th November 2008]
- House, D. (1990). *Tonal Perception in Speech*. Lund: Lund University Press.
- Hove, I. (2002). *Die Aussprache der Standardsprache in der deutschen Schweiz*. Tübingen: Max Niemeyer.
- Howell, D. C. (2007). *Statistical Methods for Psychology* (6th ed.). Belmont, CA: Thomson Wadsworth.
- Howell, P. (1984). *An acoustic determinant of perceived and produced isochrony*. Paper presented at the 10th International Congress of Phonetic Sciences Utrecht, The Netherlands, 1st-6th August 1984.

- Howell, P. (1988). Prediction of P-centre location from the distribution of energy in the amplitude envelope: I. *Perception and Psychophysics*, 43, 90-93.
- Hume, E., & Johnson, K. (Eds.). (2001). *The Role of Speech Perception in Phonology*. San Diego and London: Academic Press.
- Hyman, L. (1975). *Phonology: Theory and Analysis*. New York: Holt, Chicago: Rinehart and Winston.
- Ingram, D. (1989). *First Language Acquisition: Method, description and explanation*. Cambridge: Cambridge University Press.
- Isačenko, A. v., & Schädlich, H.-J. (1966). Untersuchungen über die deutsche Satzintonation. *Studia Grammatica*, 7, 7-67. Berlin: Akademie Verlag.
- Iversen, J. R., Patel, A. D., & Ohgushi, K. (2009). Perception of rhythmic grouping depends on auditory experience. *Journal of the Acoustical Society of America*, 124(4), 2263-2271.
- Jakobson, R., Fant, G., & Halle, M. (1952). Preliminaries to speech analysis: The Distinctive Features and their correlates. *Acoustics Laboratory, MIT, Technical Report*, 13.
- Jankowski, L. (2001). Replicating the Speech Cycling Task paradigm with French material. In C. Cavé, I. Guaitella & S. Santi (Eds.), *Actes du colloque ORAGE 2001, ORAlité et GEstualité – Interactions et comportements multimodaux dans la communication, Aix-en-Provence, 18th-22nd June 2001* (pp. 610-614).
- Jassem, W., Hill, D. R., & Witten, I. H. (1984). Isochrony in English speech: its statistical validity and linguistic relevance. In D. Gibbon & H. Richter (Eds.), *Intonation, Accent, and Rhythm* (pp. 203-225). Berlin and New York: Walter de Gruyter.
- Jassem, W., Morton, J., & Steffen-Batóg, M. (1968). The perception of stress in synthetic speech-like stimuli by Polish listeners. *Speech Analysis and Synthesis*, 1, 289-308.
- Jeon, H.-S. (2006). *Acoustic Measure of Speech Rhythm: Korean Learners of English*. Unpublished MSc thesis, University of Edinburgh, Edinburgh.
- Jespersen, O. (1933). *Linguistica, Selected Papers in English, French, and German*. Copenhagen: Levin and Munksgaard.
- Jessen, M., Marasek, K., Schneider, K., & Classen, K. (1995). *Acoustic correlates of word stress and the tense/lax opposition in the vowel system of German*. Paper presented at the 13th International Congress of Phonetic Sciences, Stockholm, Sweden, 13th-19th August 1995.
- Jinbo, K. (1980). Kokugo no onseijou no tokushitsu (The top phonetic characteristics of Japanese). In Shibata, Kitamura & Kindaichi (Eds.), *Nibon no gengogaku (Linguistics of Japan)* (pp. 5-15). Tokyo: Taishukan.

- Jomori, I., & Hoshiyama, M. (2009). Auditory brain response modified by temporal deviation of language rhythm: An auditory event-related potential study. *Neuroscience Research*, *65*, 187-193.
- Jones, D. (1956). *An outline of English Phonetics* (8th ed.). Cambridge: Heffer.
- Jun, S.-A. (1998). The Accentual Phrase in the Korean prosodic hierarchy. *Phonology*, *15*, 189-226.
- Jun, S.-A., & Fougeron, C. (1995). *The Accentual Phrase and the Prosodic Structure of French*. Paper presented at the 13th International Congress of Phonetic Sciences, Stockholm, Sweden, 13th-19th August 1995.
- Jun, S.-A., & Fougeron, C. (2000). A phonological model of French intonation. In A. Botinis (Ed.), *Intonation: Analysis, Modelling and Technology* (pp. 209-242). Dordrecht and London: Kluwer Academic.
- Jusczyk, P. W. (1997). *The Discovery of Spoken Language*. Cambridge, MA. and London: MIT Press.
- Jusczyk, P. W., & Thompson, E. (1978). Perception of a phonetic contrast in multisyllabic utterances by two month olds. *Perception and Psychophysics*, *2*, 105 -109.
- Kamiyama, T. (2003). *L'allongement en fin de phrases lues en français – une étude sur des phrases courtes chez les locuteurs natifs et les apprenants*. Paper presented at 'Sixièmes Rencontres Jeunes Chercheurs de l'Ecole Doctorale 268 "Langage et Langues"', Paris, France, May 2003.
- Keane, E. (2006). Rhythmic characteristics of colloquial and formal Tamil. *Language and Speech*, *49*(3), 299-332.
- Keller, E., Bailly, G., Monaghan, A., Terken, J., & Huckvale, M. (Eds.). (2002). *Improvements in Speech Synthesis*. Chichester: John Wiley & Sons.
- Keller, R. E. (1961). *German Dialects: Phonology and Morphology; with selected texts*. Manchester: Manchester University Press.
- Kelly, J. (1993). David Abercrombie (Obituary). *Phonetica*, *50*, 68-71.
- Kenning, M.-M. (1979). Intonation systems in French. *Journal of the International Phonetic Association*, *9*, 15-30.
- Kim, J., Davis, C., & Cutler, A. (2008). Perceptual tests of rhythmic similarity: II. Syllable rhythm. *Language and Speech*, *51*(4), 343-359.
- Kingdon, R. (1958). *The Groundwork of English Intonation*. London: Longman.
- Kingston, J., Kawaharab, S., Chambless, D., Masha, D., & Brenner-Alsopa, E. (2009). Contextual effects on the perception of duration. *Journal of Phonetics*, *37*, 297-320.

- Knecht, P. (1979). Le français en Suisse romande: Aspects linguistiques et sociolinguistiques. In A. Valdman (Ed.), *Le français hors de France* (pp. 249-258). Paris: Honoré Champion.
- Knecht, P. (1982). Die französischsprachige Schweiz. In R. Schläpfer (Ed.), *Die vier-sprachige Schweiz* (pp. 163-209). Zürich and Cologne: Benziger.
- Knecht, P., & Rubattel, C. (1984). A propos de la dimension sociolinguistique du français en Suisse romande. *Le français moderne*, 52(3/4), 138-150.
- Kochanski, G., Grabe, E., Coleman, J., & Rosner, B. (2005). Loudness predicts prominence: Fundamental frequency lends little. *Journal of the Acoustical Society of America*, 118(2), 1038-1054.
- Kohler, K. J. (2000). Investigating unscripted speech: implications for phonetics and phonology. *Phonetica*, 57, 85-94.
- Kohler, K. J. (2008). The perception of prominence patterns. *Phonetica*, 65, 257-269.
- Kohler, K. J. (2009a). Rhythm in speech and language: A new research paradigm. *Phonetica*, 66, 29-45.
- Kohler, K. J. (2009b). Whither speech rhythm research? *Phonetica*, 66, 5-14.
- Kolinsky, R., Cuvelier, H., Goetry, V., Peretz, I., & Morais, J. (2009). Music training facilitates lexical stress processing. *Music Perception*, 26(3), 235-246.
- Kozhevnikov, V. A., & Chistovic, L. S. (1965). *Speech: Articulation and Perception*. Washington: U.S. Department of Commerce, Clearinghouse for Federal Scientific and Technical Information, Joint Publications Research Service.
- Krull, D., & Engstrand, O. (2003). Speech rhythm – intention or consequence? Cross-language observation on the hyper/hypo dimension. *PHONUM*, 9, 133-136.
- Kügler, F. (2004). The phonology and phonetics of nuclear rises in Swabian German. In P. Gilles & J. Peters (Eds.), *Regional Variation in Intonation* (pp. 75-98). Tübingen: Niemeyer.
- Kuhl, P., & Iverson, P. (1995). Linguistic experience and the “Perceptual Magnet Effect”. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp. 121-154). Baltimore: York Press.
- Lacheret-Dujour, A., & Beaugendre, F. (1999). *La prosodie du français*. Paris: CNRS Editions.
- Ladd, D. R. (1996). *Intonational phonology*. Cambridge: Cambridge University Press.
- Ladefoged, P. (1967). *Three Areas of Experimental Phonetics*. London: Oxford University Press.
- Ladefoged, P. (2001). *A Course in Phonetics* (4th ed.). Boston: Heinle and Heinle.

- Ladefoged, P., Draper, M. H., & Whitteridge, D. (1958). Syllables and stress. *Miscellanea Phonetica*, 3, 1-14.
- Laver, J. (1994). *Principles of Phonetics*. Cambridge: Cambridge University Press.
- Lee, C. S., & Todd, N. P. M. (2004). Towards an auditory account of speech rhythm: application of a model of the auditory 'primal sketch' to two multi-language corpora. *Cognition*, 93, 225-254.
- Leeman, A., & Siebenhaar, B. (2007). *Intonational and Temporal Features of Swiss German*. Paper presented at the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany, 6th-10th August 2007.
- Lehiste, I. (1970). *Suprasegmentals*. Cambridge, MA and London: MIT Press.
- Lehiste, I. (1973). Rhythmic units and syntactic units in production and perception. *Journal of the Acoustical Society of America*, 54, 1228-1234.
- Lehiste, I. (1976). Influence of fundamental frequency pattern on the perception of duration. *Journal of Phonetics*, 4, 113-117.
- Lehiste, I. (1977). Isochrony reconsidered. *Journal of Phonetics*, 5, 253-263.
- Lehiste, I. (1980). Phonetic manifestation of syntactic structure in English. *Annual Bulletin of the Research Institute of Logopedics and Phoniatrics at the University of Tokyo*, 14, 1-27.
- Lehnert-LeHouillier, H. (2007). *The influence of dynamic F0 on the perception of vowel duration: cross-linguistic evidence*. Paper presented at the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany, 6th-10th August 2007.
- Léon, M. (1964). *Exercices systématiques de prononciation française*. Paris: Hachette-Larousse.
- Levitt, A. G., & Wang, Q. (1991). Evidence for language-specific rhythmic influences in the reduplicative babbling of French- and English-learning infants. *Language and Speech*, 34, 235-249.
- Liberman, A. M., Harris, K. S., Eimas, P. D., Lisker, L., & Bastian, J. (1961). An effect of learning on speech perception: The discrimination of durations of silence with and without phonemic significance. *Language and Speech*, 54, 175-195.
- Liberman, M. Y. (1975). *The Intonational System of English*. Unpublished doctoral thesis, MIT.
- Liberman, M. Y., & Prince, A. (1977). On stress and linguistic rhythm. *Linguistic Enquiry*, 8, 249-336.
- Lieberman, P. (1960). Some acoustic correlates of word stress in American English. *Journal of the Acoustical Society of America*, 32(4), 451-454.

- Lievano, S. J., & Egger, N. (2005). *Hoi: Your Swiss German Survival Guide*. Basel: Bergli Books.
- Lisker, L., & Abramson, A. S. (1970). *The voicing dimension: some experiments in comparative phonetics*. Paper presented at the 6th International Congress of Phonetic Sciences, Prague, Czech Republic 1970.
- Liss, J. M., White, L., Mattys, S., Lansford, K., Lotto, A. J., Spitzer, S. M., et al. (2009). Quantifying speech rhythm abnormalities in the dysarthrias. *Journal of Speech, Language and Hearing Research*, 52, 1334-1352.
- Lleó, C., Rakow, M., & Kehoe, M. (2007). *Acquiring Rhythmically Different Languages in a Bilingual Context*. Paper presented at the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany, 6th-10th August 2007.
- Llisterri, J., Machuca, M., de la Mota, C., Riera, M., & Rios, A. (2003). *The perception of lexical stress in Spanish*. Paper presented at the 15th International Congress of Phonetic Sciences, Barcelona, Spain, 3rd-9th August 2003.
- Lloyd James, A. (1940). *Speech Signals in Telephony*. London: Sir Isaac Pitman & Sons.
- Lodge, R. A. (1993). *French: from dialect to standard*. London: Routledge.
- Logan, J. S., & Pruitt, J. S. (1995). Methodological issues in training listeners to perceive non-native phonemes. In W. Strange (Ed.), *Speech Perception and Linguistic Experience: Issues in Cross-language Research* (pp. 351-377). Baltimore: York Press.
- Low, E. L. (1994). *Intonation Patterns in Singapore English*. Unpublished MPhil thesis, University of Cambridge, Cambridge.
- Low, E. L. (1998). *Prosodic Prominence in Singapore English*. Unpublished doctoral thesis, University of Cambridge, Cambridge.
- Low, E. L., & Grabe, E. (1995). *Prosodic Patterns in Singapore English*. Paper presented at the 13th International Congress of Phonetic Sciences, Stockholm, Sweden, 13th-19th August 1995.
- Low, E. L., Grabe, E., & Nolan, F. (2000). Quantitative characterizations of speech rhythm: Syllable-timing in Singapore English. *Language and Speech*, 43(4), 377-401.
- Lucas, D. W. (1968). *Aristotle's Poetics*. Oxford: Clarendon Press.
- Lüdi, G., & Werlen, I. (2005). *Le paysage linguistique en Suisse*. Neuchâtel: Swiss Federal Statistics Office.
- Lyche, C., & Girard, F. (1995). Le mot retrouvé. *Lingua*, 95(1-3), 205-221.
- Maas, U. (1999). *Phonologie: Einführung in die funktionale Phonetik des Deutschen*. Opladen, Wiesbaden: Westdeutscher Verlag.

- MacCarthy, P. (1975). *The Pronunciation of German*. London: Oxford University Press.
- Macmillan, N. A., & Creelman, C. D. (1991). *Detection Theory: a user's guide*. Cambridge: Cambridge University Press.
- Magne, C., Aramaki, M., Astesano, C., Gordon, R. L., Ystad, S., Farner, S., et al. (2004). Comparison of rhythmic processing in language and music – an interdisciplinary approach. *The Journal of Music and Meaning* 3, section 5.
- Manno, G. (2004). Le français régional de Suisse romande. In A. Coveney, M.-A. Hintze & C. Sanders (Eds.), *Variation et francophonie* (pp. 331-357). Paris: L'Harmattan.
- Marcus, S. M. (1981). Acoustic determinants of perceptual centre (P-centre) location. *Perception and Psychophysics*, 30, 247-256.
- Marouzeau, J. (1924). Accent affectif et accent intellectuel. *Bulletin de la Société de Linguistique de Paris*, 25, 79-86.
- Marques, C., Moreno, S., Castro, S. L., & Besson, M. (2007). Musicians detect pitch violation in a foreign language better than nonmusicians: behavioral and electrophysiological evidence. *Journal of Cognitive Neuroscience*, 19(9), 1453-1463.
- Martin, J. G. (1972). Rhythmic (hierarchical) versus serial structure in speech and other behavior. *Psychological Review* 79(6), 487-509.
- Martin, P. (1980). Une théorie syntaxique de l'accentuation en français. In P. R. Léon & M. Rossi (Eds.), *L'accent en français contemporain: Studia Phonetica*, 15 (pp. 1-12). Montréal: Didier.
- Martin, P. (1987). Prosodic and rhythmic structures in French. *Linguistics*, 25(5), 925-950.
- Mattingly, I. G., Liberman, A. M., Syrdal, A. M., & Halwes, T. (1971). Discrimination in speech and nonspeech modes. *Cognitive Psychology*, 2, 131-157.
- McAuley, J. D. (1995). *Perception of time as phase: toward an adaptive-oscillator model of rhythmic pattern processing*. Unpublished doctoral thesis, Indiana University.
- McLennan, S. (2005). *Linguistic Rhythm: A Dynamical Model*. Poster presented at PSP2005, London, UK, 16th June 2005.
- McLennan, S., & Hockema, S. (2002). Spike-v: An adaptive mechanism for speech-rate independent timing. *Indiana University Working Papers Online*, 2.
- Mehler, J. (1981). The role of syllables in speech processing: Infant and adult data. *Philosophical Transactions of the Royal Society*, 295, 333-352.
- Mehler, J., Jusczyk, P., Lambertz, G., Halsted, N., Bertocini, J., & Amiel-Tison, C. (1988). A precursor of language acquisition in young infants. *Cognition*, 29, 143-178.

- Mehlhorn, G. (2007). Individual pronunciation coaching and prosody. In J. Trouvain & U. Gut (Eds.), *Non-native Prosody: Phonetic Description and Teaching Practice* (pp. 211-236). Berlin: Mouton de Gruyter.
- Menard, S. (2004). Six approaches to calculating standardized logistic regression coefficients. *The American Statistician*, 58, 218-223.
- Mertens, P. (1987). *L'intonation du français. De la description linguistique à la reconnaissance automatique*. Unpublished doctoral dissertation, Katholieke Universiteit Leuven.
- Mertens, P. (1991). *Local prominence of acoustic and psychoacoustic functions and perceived stress in French*. Paper presented at the 12th International Congress of Phonetic Sciences, Aix-en-Provence, France, 19th-24th August 1991.
- Mertens, P. (1992). L'accentuation de syllabes contiguës. *ITL Review of Applied Linguistics*, 95-96, 145-165.
- Mertens, P. (1993). Accentuation, intonation et morphosyntaxe. *Travaux de Linguistique*, 26, 21-69.
- Métral, J.-P. (1977). Le vocalisme du français en Suisse romande. *Cahiers Ferdinand de Saussure*, 31, 145-176.
- Miller, J. S. (2007). *Swiss French Prosody: Intonation, rate and speaking style in the Vaud canton*. Unpublished doctoral thesis, University of Illinois at Urbana-Champaign, Urbana, Illinois.
- Miller, M. (1984). On the perception of rhythm. *Journal of Phonetics*, 12, 75-83.
- Missaglia, F. (2007). Prosodic training of Italian learners of German: the Contrastive Prosody Method. In J. Trouvain & U. Gut (Eds.), *Non-native Prosody: Phonetic Description and Teaching Practice* (pp. 237-258). Berlin: Mouton de Gruyter.
- Mixdorff, H. (1998). *Intonation Patterns of German*. Unpublished doctoral thesis, Technische Universität, Dresden.
- Miyake, I. (1902). Researches on rhythmic action. *Studies from the Yale Psychological Laboratory*, 10, 2-48.
- Monaghan, A. (2002). Prosody in synthetic speech: Problems, solutions and challenges. In E. Keller, G. Bailly, A. Monaghan, J. Terken & M. Huckvale (Eds.), *Improvements in Speech Synthesis* (pp. 89-92). Chichester: John Wiley & Sons.
- Moore, B. C. J. (2004). *An Introduction to the Psychology of Hearing* (5th ed.). London and San Diego: Elsevier Academic Press.
- Morrongiello, B. A. (1984). Auditory temporal pattern perception in 6- and 12-month-old infants. *Developmental Psychology*, 20, 441-448.

- Morton, J., & Jassem, W. (1965). Acoustic correlates of stress. *Language and Speech*, 8, 159-181.
- Morton, J., Marcus, S., & Frankish, C. (1976). Perceptual Centers. *Psychological Review*, 83(5), 405-408.
- Moulines, E., & Charpentier, F. (1990). Pitch-synchronous waveform processing techniques for text-to-speech synthesis using diphones. *Speech Communication*, 9, 453-467.
- Munro, M. J. (1995). Nonsegmental factors in foreign accent: ratings of filtered speech. *Studies in Second Language Acquisition*, 17, 17-34.
- Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Article reprinted in Language Learning*, 49(Supplement 1), 285-310. [Originally published as Munro, M. J., & Derwing, T. (1995). *Language Learning*, 1945, 1973-1997.]
- Murty, L., Otake, T., & Cutler, A. (2007). Perceptual tests of rhythmic similarity: I. Mora rhythm. *Language and Speech*, 50, 77-99.
- Navarro Tomas, T. (1916). Cantidad de las vocales acentuadas. *Revista de Filología Española*, III, 387-408.
- Navarro Tomas, T. (1917). Cantidad de las vocales inacentuadas. *Revista de Filología Española*, IV, 371-388.
- Nazzi, T. (1997). *Du rythme dans l'acquisition et le traitement de la parole [Rhythm in the acquisition and the processing of speech]*. Unpublished doctoral thesis, Ecole des Hautes Etudes en Sciences Sociales, Paris.
- Nazzi, T., & Ramus, F. (2003). Perception and acquisition of linguistic rhythm by infants. *Speech Communication*, 41, 233-243.
- Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance*, 24(3), 756-766.
- Nazzi, T., Iakimova, G., Bertoncini, J., Fredonie, S., & Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences. *Journal of Memory and Language*, 54(3), 283-299.
- Nazzi, T., Jusczyk, P. W., & Johnson, E. K. (2000). Language discrimination by English-learning 5-month-olds: Effects of rhythm and familiarity. *Journal of Memory and Language*, 43, 1-19.
- Nespor, M., & Vogel, I. (1986). *Prosodic phonology*. Dordrecht and Riverton, N.J.: Foris.
- Niebuhr, O. (2008). *The effect of global rhythms on local accent perceptions in German*. Paper presented at Speech Prosody 2008, Campinas, Brazil, 6th-9th May 2008.

- Niebuhr, O. (2009). F0-based rhythm effects on the perception of local syllable prominence. *Phonetica*, 66, 95-112.
- Nolan, F. (2003). *Intonational equivalence: an experimental evaluation of pitch scales*. Paper presented at the 15th International Congress of Phonetics Sciences, Barcelona, Spain, 3rd-9th August 2003.
- Nolan, F., & Asu, E. L. (2009). The Pairwise Variability Index and coexisting rhythms in language. *Phonetica*, 66, 64-77.
- Nolan, F., & Hausmann, T. (2005). *Are phrase tones in Swiss German intonation stress-seeking?* Poster presented at the Between Stress and Tone Conference, Leiden, The Netherlands, 16th-18th June 2005.
- Nübling, D., & Schrambke, R. (2004). Silben- versus akzentsprachliche Züge in germanischen Sprachen und im Alemannischen. In E. Glaser, P. Ott & R. Schwarzenbach (Eds.), *Alemannisch im Sprachvergleich. Beiträge zur 14. Arbeitstagung für alemannische Dialektologie in Männedorf, Zürich* (pp. 281-320). Wiesbaden: Franz Steiner Verlag.
- Nyrop, K. (1963). *Manuel de phonétique du français parlé* (8th ed.). New York: Stechert.
- O'Connor, J. D. (1965). The perception of time intervals. *Progress Report, Department of Phonetics, University College London*, 2, 11-15.
- O'Connor, J. D. (1973). *Phonetics*. Harmondsworth: Penguin.
- Offord, M. H. (1990). *Varieties of Contemporary French*. Basingstoke: Macmillan Education.
- Ohala, J. J. (1975). The temporal regulation of speech. In G. Fant & M. A. A. Tatham (Eds.), *Auditory Analysis and Perception of Speech* (pp. 431-453). London and New York: Academic Press.
- Ohala, J. J. (1978). Production of tone. In V. A. Fromkin (Ed.), *Tone: A Linguistic Survey* (pp. 5-39). New York: Academic Press.
- Oxford English Dictionary*. (1989). (2nd ed.) Oxford: Oxford University Press. Accessed online at <<http://www.oed.com>>
- Palmer, H. (1922). *English Intonation, with systematic exercises*. Cambridge: Heffer.
- Pamies Bertrán, A. (1999). Prosodic typology: On the dichotomy between stress-timed and syllable-timed languages. *Language Design*, 2, 103-130.
- Parmenter, C. E., & Blanc, A. V. (1933). An experimental study of accent in French and English. *PMLA*, 48, 598-607.
- Partridge, E. (1961). *A short etymological dictionary of modern English*. London: Routledge and Kegan Paul.

- Pasquini, E. S., Corriveau, K. H., & Goswami, U. (2007). Auditory processing of amplitude envelope rise time in adults diagnosed with developmental dyslexia. *Scientific Studies of Reading, 11*(3), 259-286.
- Patel, A. D. (2007). *Music, Language and the Brain*. Oxford: Oxford University Press.
- Patel, A. D., & Daniele, J. R. (2003). An empirical comparison of rhythm in language and music. *Cognition, 87*, B35-B45.
- Patel, A. D., & Iversen, J. R. (2007). The linguistic benefits of musical abilities. *Trends in Cognitive Sciences, 11*(9), 369-372.
- Patel, A. D., Iversen, J. R., & Rosenberg, J. C. (2006). Comparing the rhythm and melody of speech and music: The case of British English and French. *Journal of the Acoustical Society of America, 119*, 3034-3047.
- Patel, A. D., Löfqvist, A., & Naito, W. (1999). *The acoustics and kinematics of regularly timed speech: a database and method for the study of the P-center problem*. Paper presented at the 14th International Congress of Phonetic Sciences, San Francisco, USA, 1st-7th August 1999.
- Payne, E., Post, B., Astruc, L., Prieto, P., & del Mar Vanrell, M. (2009). Rhythmic modification in child directed speech. *Oxford University Working Papers in Linguistics, Philology & Phonetics, 12*, 123-144.
- Peperkamp, S., & Dupoux, E. (2002). A typological study of stress 'deafness'. In C. Gussenhoven & N. Warner (Eds.), *Papers in Laboratory Phonology VII* (pp. 203-240). Berlin: Mouton de Gruyter.
- Perey, A. J., & Pisoni, D. B. (1980). Identification and discrimination of durations of silence in nonspeech signals. *Research on Speech Perception: Department of Psychology, Indiana University, Progress Report, 6*, 235-269.
- Peterson, G. E., & Lehiste, I. (1960). Duration of syllable nuclei in English. *Journal of the Acoustical Society of America, 32*(6), 693-703.
- Pierrehumbert, J. (1980). *The Phonology and Phonetics of English Intonation*. Doctoral thesis, MIT, Cambridge, Mass.
- Pike, K. (1945). *The Intonation of American English*. Ann Arbor: University of Michigan Press.
- Pinker, S., & Jackendoff, R. (2005). The faculty of language: what's special about it? *Cognition, 95*, 201-236.
- Piske, T., MacKay, I. R. A., & Flege, J. E. (2001). Factors affecting degree of foreign accent in an L2: a review. *Journal of Phonetics, 29*, 191-215.

- Pisoni, D. B. (1976). Fundamental frequency and perceived vowel duration. *Journal of the Acoustical Society of America*, 59(S1), S39.
- Pointon, G. (1980). Is Spanish really syllable-timed? *Journal of Phonetics*, 8, 293-304.
- Pollack, I., & Pisoni, D. B. (1971). On the comparison between identification and discrimination tests in speech perception. *Psychological Science*, 24(6), 299-300.
- Pompino-Marschall, B. (1989). On the psychoacoustic nature of the P-centre phenomenon. *Journal of Phonetics*, 17, 175-192.
- Port, R. F., Cummins, F., & Gasser, M. (1995). A dynamic approach to rhythm in language: Toward a Temporal Phonology. In B. Luka & B. Need (Eds.), *Proceedings of the Chicago Linguistics Society* (pp. 375-397). Department of Linguistics: University of Chicago.
- Post, B. (2000). *Tonal and Phrasal Structures in French Intonation*. Doctoral thesis, Catholic University of Nijmegen, Nijmegen.
- Post, B. (2002). *French Tonal Structures*. Paper presented at Speech Prosody 2002, Aix-en-Provence, France, 11th-13th April 2002.
- Prince, A. (1983). Relating to the grid. *Linguistic Enquiry*, 14, 19-100.
- Pulgram, E. (1965). Prosodic systems: French. *Lingua*, 13, 125-144.
- Pulvermüller, F. (2002). *The Neuroscience of Language: On Brain Circuits of Words and Serial Order*. Cambridge: Cambridge University Press.
- Ramus, F. (2002). *Acoustic Correlates of Linguistic Rhythm: Perspectives*. Paper presented at Speech Prosody 2002, Aix-en-Provence, France, 11th-13th April 2002.
- Ramus, F., & Mehler, J. (1999). Language identification with suprasegmental cues: A study based on speech synthesis. *Journal of the Acoustical Society of America*, 105(1), 512-521.
- Ramus, F., Dupoux, E., & Mehler, J. (2003). *The psychological reality of rhythm classes: perceptual studies*. Paper presented at the 15th International Congress of Phonetic Sciences, Barcelona, Spain, 3rd-9th August 2003.
- Ramus, F., Hauser, M., Miller, C., Morris, D., & Mehler, J. (2000). Language discrimination by human newborns and cotton-top tamarin monkeys. *Science*, 288, 349-351.
- Ramus, F., Nespore, M., & Mehler, J. (1999). Correlates of linguistic rhythm in the speech signal. *Cognition*, 73(3), 263-292.
- Rash, F. J. (1998). *The German Language in Switzerland: Multilingualism, Diglossia and Variation*. Bern: Peter Lang.

- Reese, J. (2007). *Swiss German: The Modern Alemannic Vernacular in and around Zurich*. Munich: Lincom Europa.
- Reeves, C., Schmauder, A. R., & Morris, R. K. (2000). Stress grouping improves performance on an immediate serial list recall task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 26(6), 1638-1654.
- Repp, B. H. (1982). Phonetic trading relations and context effects: New experimental evidence for a speech mode of perception. *Psychological Bulletin*, 92(1), 81-110.
- Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), *Speech and Language: Advances in Basic Research and Practice* (Vol. 10, pp. 243-335). Orlando FL: Academic Press.
- Richardson, U., Thomson, J., Scott, S. K., & Goswami, U. (2004). Auditory processing skills and phonological representation in dyslexic children. *Dyslexia: An International Journal of Research & Practice*, 10, 215-233.
- Rigault, A. (1962). *Rôle de la fréquence, de l'intensité et de la durée vocalique dans la perception de l'accent en français*. Paper presented at the 4th International Congress of Phonetic Sciences, Helsinki, Finland, 4th-9th September 1962.
- Roach, P. (1982). On the distinction between 'stress-timed' and 'syllable-timed' languages. In D. Crystal (Ed.), *Linguistic Controversies: Essays in honour of F.R. Palmer* (pp. 73-79). London: Edward Arnold.
- Rosen, S. M. (1977a). Fundamental frequency patterns and the long-short vowel distinction in Swedish. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 1, 31-37.
- Rosen, S. M. (1977b). The Effect of fundamental frequency patterns on perceived duration. *Speech Transmission Laboratory Quarterly Progress and Status Report*, 1, 17-30.
- Rossi, M. (1971). Le seuil de glissando ou seuil de perception des variations tonales pour les sons de la parole. *Phonetica*, 23, 1-33.
- Rossi, M. (1978). La perception des glissandos descendants dans les contours prosodiques. *Phonetica*, 35, 11-40.
- Rossi, M. (1980). Le français, langue sans accent? In P. R. Léon & M. Rossi (Eds.), *L'accent en Français Contemporain, Studia Phonetica*, 15 (pp. 13-51). Montréal: Didier.
- Rousselot, P.-J. (1924). *Principes de phonétique expérimentale*. Paris: Didier.
- Salmelin, R., Schnitzler, A., Parkkonen, L., Biermann, K., Helenius, P., Kiviniemi, K., et al. (1999). Native language, gender, and functional organization of the auditory cortex. *Proceedings of the National Academy of Sciences*, 96, 10460-10465.

- Sansavini, A. (1997). Neonatal perception of the rhythmical structure of speech: the role of stress patterns. *Early Development and Parenting*, 6(1), 3-13.
- Sautermeister, P., & Eklund, R. (1997). *Some Observations on the Influence of F0 and Duration to the Perception of Prominence by Swedish Listeners*. Paper presented at FONETIK 1997, Umeå, Lövånger, Sweden 1997.
- Schiering, R. (2006). *Towards a Typology of Linguistic Rhythm*. Paper presented at the Manchester Phonology Meeting, Manchester, UK 2006.
- Schiering, R. (2007). The phonological basis of linguistic rhythm: cross-linguistic data and diachronic interpretation. *Sprachtypologie und Universalienforschung*, 60(4), 337-359.
- Schläpfer, R. (1982). *Die vier-sprachige Schweiz*. Zürich and Cologne: Benziger.
- Schlüter, J. (2005). *Rhythmic Grammar: the influence of rhythm on grammatical variation and change in English*. Berlin: Walter de Gruyter.
- Schmid, S. (2001). *Un nouveau fondement phonétique pour la typologie rythmique des langues?* Poster presented at the workshop for 'le 10ème anniversaire du Laboratoire d'analyse informatique de la parole (LAIP)', Lausanne, Switzerland, 31st May 2001.
- Schmid, S. (2004). Une approche phonétique de l'isochronie dans quelques dialectes italo-romans. In T. Meisenburg & M. Selig (Eds.), *Nouveaux départs en phonologie* (pp. 109-124). Tübingen: G. Narr.
- Schneider, K., & Möbius, B. (2007). *Word stress correlates in spontaneous child-directed speech in German*. Paper presented at Interspeech 2007, Antwerp, Belgium, 27th-31st August 2007.
- Schoch, M. (1978). Problème sociolinguistique des pronoms d'allocution: 'Tu' et 'Vous'. Enquête à Lausanne. *La Linguistique*, 14, 55-73.
- Scholes, R. J. (1971). On the spoken disambiguation of superficially ambiguous sentences. *Language and Speech*, 14, 1-11.
- Schön, D., Magne, C., & Besson, M. (2004). The music of speech: Music training facilitates pitch processing in both music and language. *Psychophysiology*, 41, 341-349.
- Scott, D. R., Isard, S. D., & de Boysson-Bardies, B. (1985). Perceptual isochrony in English and French. *Journal of Phonetics*, 13, 155-162.
- Scott, D. R., Isard, S. D., & de Boysson-Bardies, B. (1986). On the measurement of rhythmic irregularity: a reply to Benguerel. *Journal of Phonetics*, 14, 327-330.
- Scott, S. K. (1994). *P-Centres in speech: an acoustic analysis*. Unpublished doctoral thesis, University College London, London.
- Scott, S. K. (1998). The point of P-centres. *Psychological Research*, 61, 4-11.

- Scott, S. K., Clegg, F., Rudge, P., & Burgess, P. (2006). Foreign accent syndrome, speech rhythm and the functional neuronatomy of speech production. *Journal of Neurolinguistics*, 19(5), 370-384.
- Scripture, E. W. (1897). *The New Psychology*. New York: Charles Scribner's Sons.
- Scripture, E. W. (1899). Researches in experimental phonetics. Observations on rhythmic action. *Studies from the Yale Psychological Laboratory*, VII.
- Scripture, E. W. (1902). *The Elements of Experimental Phonetics*. New York: Charles Scribner's Sons.
- Sebastián-Gallés, N., Dupoux, E., Segui, J., & Mehler, J. (1992). Contrasting syllabic effects in Catalan and Spanish. *Journal of Memory and Language*, 31, 18-32.
- Selkirk, E. (1984). *Phonology and syntax: The relation between sound and structure*. Cambridge, MA: MIT Press.
- Sergeant, R. L., & Harris, J. D. (1962). Sensitivity to unidirectional frequency modulation. *Journal of the Acoustical Society of America*, 34, 1625-1628.
- Shattuck-Hufnagel, S., & Turk, A. E. (1996). A prosody tutorial for investigators of auditory sentence processing. *Journal of Psycholinguistic Research*, 2, 193-247.
- Shen, Y., & Peterson, G. G. (1962). Isochronism in English. *Studies in Linguistics: University of Buffalo Occasional Papers*, 9, 1-36.
- Siebenhaar, B. (2004). Comparing timing models of two Swiss German dialects. In B.-L. Gunnarsson, L. Bergström, G. Eklund & S. Fridell (Eds.), *Language Variation in Europe. Papers from ICLaVE 2* (pp. 353-365). Uppsala.
- Siebenhaar, B. (2005). Die Modellierung zeitlicher Strukturen im Schweizerdeutschen. In E. Eggers, J. E. Schmidt & D. Stellmacher (Eds.), *Moderne Dialekte – Neue Dialektologie* (pp. 343-361). Stuttgart: Steiner.
- Siebenhaar, B. (2006). Das sprachliche Normenverständnis in mundartlichen Chaträumen der Schweiz. In J. Androutsopoulos, J. Runkehl, P. Schlobinski & T. Siever (Eds.), *Neuere Entwicklungen in der linguistischen Internetforschung* (pp. 45-67). Zürich and New York: Hildesheim.
- Siebenhaar, B., & Vögeli, W. (1997). Mundart und Hochdeutsch im Vergleich. Revision of chapter for second edition of P. Sieber & H. Sitta (Eds.), *Mundart und Hochdeutsch im Unterricht. Orientierungshilfen für Lehrer*. Aarau, Frankfurt am Main, Salzburg: Studienbücher Sprachlandschaft (unpublished).

- Siebenhaar, B., Forst, M., & Keller, E. (2004). Prosody of Bernese and Zurich German. What the development of a dialectal speech synthesis system tells us about it. In P. Gilles & J. Peters (Eds.), *Regional Variation in Intonation* (pp. 219-238). Tübingen: Niemeyer.
- Siebenhaar, B., Forst, M., & Keller, E. (2006). Speech synthesis of dialectal variants as a method for research on prosody. In M. Filippula, J. Klemola, M. Palander & E. Penttilä (Eds.), *Topics in Dialectal Variation* (pp. 145-162). Joensuu: Joensuun yliopistopaino.
- Sievers, E. (1912). *Rhythmisch-melodische Studien*. Heidelberg: Winter.
- Simon, A. C. (2003). La variation prosodique régionale en français dans les données conversationnelles: propositions théoriques et méthodologiques. *Bulletin Phonologie du Français Contemporain*, 3, 99-113.
- Singy, P. (1996). *L'image du français en Suisse romande: Une enquête sociolinguistique en Pays de Vaud*. Paris and Montreal: L'Harmattan.
- Singy, P. (2001). Exterritorialité de la norme linguistique de prestige et représentations linguistiques: les disparités entre générations en Suisse romande. In T. Pooley & E. Minz (Eds.), *French Accents* (pp. 269-287). London: CILT.
- Sluijter, A. M. C., & van Heuven, V. J. (1996). Spectral balance as an acoustic correlate of linguistic stress. *Journal of the Acoustical Society of America*, 100(4), 2471-2485.
- Sluijter, A. M. C., van Heuven, V. J., & Pacilly, J. A. (1997). Spectral balance as a cue in the perception of linguistic stress. *Journal of the Acoustical Society of America*, 101(1), 503-513.
- Smith, C. L. (2002). Prosodic finality and sentence type in French. *Language and Speech* 45(2), 141-178.
- Smith, M. R., Cutler, A., Butterfield, S., & Nimmo-Smith, I. (1989). The perception of rhythm and word boundaries in noise-masked speech. *Journal of Speech and Hearing Research*, 32, 912-920.
- Spitznagel, A. (2000). Zur Geschichte der psychologischen Rhythmusforschung. In K. Müller & G. Aschersleben (Eds.), *Rhythmus: ein interdisziplinäres Handbuch* (pp. 1-40). Bern: Hans Huber Verlag.
- Sprachatlas der deutschen Schweiz*. (1962-2003). Ed. by R. Hotzenköcherler, H. Baumgartner, K. Lobeck, R. Schläpfer, R. Trüb, & P. Zinsli. Bern: Francke.
- Spring, D. R., & Dale, P. S. (1977). Discrimination of linguistic stress in early infancy. *Journal of Speech and Hearing Research*, 20, 224-232.
- Stedje, A. (1999). *Deutsche Sprache gestern und heute* (4th ed.). Munich: Wilhelm Fink.

- Steele, J. (1775). *Prosodia rationalis: Or, an essay towards establishing the melody and measure of speech*. London: J. Nichols.
- Steinberg, J. (1996). *Why Switzerland?* (2nd ed.). Cambridge: Cambridge University Press.
- Steiner, I. (2004). *Zur Rhythmusanalyse mittels akustischer Parameter*. Unpublished MA thesis, University of Bonn, Bonn.
- Steiner, I. (2005). On the analysis of speech rhythm through acoustic parameters. In B. Fisseni, H.-C. Schmitz, B. Schröder & P. Wagner (Eds.), *Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung 2005 in Bonn* (pp. 647-658). Frankfurt am Main: Peter Lang.
- Stone, M. (1979). Manifestations of rhythm and stress in physiological measurements of jaw activity in speech. *Journal of the Acoustical Society of America*, 65, S25-S25.
- Strange, W. (Ed.). (1995). *Speech Perception and Linguistic Experience: Issues in Cross-language Research*. Baltimore: York Press.
- Strangert, E. (1985). *Swedish speech rhythm in a cross-language perspective*. Umeå: Almqvist & Wiksell International.
- Stucki, K. (1921). *Schweizerdeutsch: Abriss einer Grammatik mit Laut- und Formenlehre*. Zürich: Orell Füssli.
- Sweet, H. (1875-6). Words, logic, and grammar. *Transactions of the Philological Society*, 439-446.
- Swerts, M. (1997). Prosodic features at discourse boundaries of different strength. *Journal of the Acoustical Society of America*, 101(1), 514-521.
- Swiss Federal Office of Statistics. (2007). Map of linguistic communities in Switzerland (from Stat@las Switzerland)
<http://www.bfs.admin.ch/bfs/portal/en/index/regionen/thematische_karten/02.html> [accessed 30th October 2009]
- Szczepek Reed, B. (2010). Speech rhythm across turn transitions in cross-cultural talk-in-interaction. *Journal of Pragmatics*, 42, 1037-1059.
- ‘t Hart, J., Collier, R., & Cohen, A. (1990). *A Perceptual Study of Intonation*. Cambridge: Cambridge University Press.
- Thompson, W. F., Schellenberg, E. G., & Husain, G. (2004). Decoding speech prosody: Do music lessons help? *Emotion*, 4(1), 46-64.
- Thorén, B. (2008). *The priority of temporal aspects in L2-Swedish prosody: Studies in perception and production*. Unpublished doctoral thesis, University of Stockholm, Stockholm.

- Tilsen, S., & Johnson, K. (2008). Low-frequency Fourier analysis of speech rhythm. *Journal of the Acoustical Society of America*, 124(2), EL34 - EL39.
- Tincoff, R., Hauser, M., Tsao, F., Spaepen, G., Ramus, F., & Mehler, J. (2005). The role of speech rhythm in language discrimination: further tests with a non-human primate. *Developmental Science*, 8(1), 26-35.
- Todd, N. P. M. (1994). The auditory “primal sketch”: a multiscale model of rhythmic grouping. *Journal of New Music Research*, 23, 25-70.
- Todd, N. P. M., & Brown, G. J. (1996). Visualization of rhythm, time and meter. *Artificial Intelligence Review*, 10, 253-273.
- Touati, P. (1987). *Structures prosodiques du suédois et du français*. Lund: Lund University Press.
- Tranel, B. (1987). *The Sounds of French*. Cambridge: Cambridge University Press.
- Trask, R. L. (1996). *A Dictionary of Phonetics and Phonology*. London and New York: Routledge.
- Trehub, S. E., & Thorpe, L. A. (1989). Infants' perception of rhythm: Categorization of auditory sequences by temporal structure. *Canadian Journal of Psychology*, 43, 217-229.
- Trouvain, J., & Gut, U. (Eds.). (2007). *Non-native Prosody: Phonetic Description and Teaching Practice*. Berlin: Mouton de Gruyter.
- Trubetzkoy, N. S. (1958). *Grundzüge der Phonologie*. Göttingen: Vandenhoeck & Ruprecht.
- Tuller, B., & Fowler, C. A. (1980). Some articulatory correlates of perceptual isochrony. *Perception and Psychophysics*, 27, 277-283.
- Ulbrich, C. (2002). *A Comparative Study of Intonation in Three Standard Varieties of German*. Paper presented at Speech Prosody 2002, Aix-en-Provence, France, 11th-13th April 2002.
- Ulbrich, C. (2004). A comparative study of declarative intonation in Swiss and German standard varieties. In P. Gilles & J. Peters (Eds.), *Regional Variation in Intonation* (pp. 99-122). Tübingen: Niemeyer.
- Ulbrich, C. (2006). *Interaction of Timing and Pitch in Cross-Varietal Data*. Paper presented at the 11th Australian International Conference on Speech Science and Technology, Auckland, New Zealand, 6th-8th December 2006.
- Vaissière, J. (1974). On French prosody. *Quarterly Progress Report (MIT)*, 114, 12-23.
- Vaissière, J. (1975). Further note on French prosody. *Quarterly Progress Report (MIT)*, 115, 251-261.
- Vaissière, J. (1980). La structuration acoustique de la phrase française. *Annali della Scuola Normale superiore di Pisa, Classe di lettere e filosofia, Serie III, X, 2*, 529-560.

- Vaissière, J. (1983). Language-independent prosodic features. In A. Cutler, D. R. Ladd & G. Brown (Eds.), *Prosody: Models and Measurements*. Berlin: Springer.
- Vaissière, J. (1991a). *Perceiving rhythm in French?* Paper presented at the 12th International Congress of Phonetic Sciences, Aix-en-Provence, France, 19th-24th August 1991.
- Vaissière, J. (1991b). Rhythm, accentuation and final lengthening in French. In J. Sundberg, L. Nord & R. Carlson (Eds.), *Music, Language, Speech and Brain: proceedings of an International Symposium at the Wenner-Gren Center, Stockholm, 5th-8th September 1990* (pp. 108-120).
- Vaissière, J. (1997). Langues, prosodies et syntaxe. *Traitement Automatique des Langues*, 38, 53-82.
- Vaissière, J. (2005). Perception of intonation. In D. B. Pisoni & R. E. Remez (Eds.), *Handbook of Speech Perception* (pp. 1-28). Oxford: Blackwell.
- van Dommelen, W. (1987). The contribution of speech rhythm and pitch to speaker recognition. *Language and Speech*, 30(4), 325-338.
- van Dommelen, W. (1991). *F0 and the perception of duration*. Paper presented at the 12th International Congress of Phonetic Sciences, Aix-en-Provence, France, 19th-24th August 1991.
- van Dommelen, W. (1993). Does dynamic F0 increase perceived duration? New light on an old issue. *Journal of Phonetics*, 21, 367-386.
- van Santen, J. P. H. (2005). Phonetic knowledge in text-to-speech Synthesis. In W. J. Barry & W. van Dommelen (Eds.), *The Integration of Phonetic Knowledge in Speech Technology* (pp. 149-166). Dordrecht: Springer.
- Vanderslice, R., & Ladefoged, P. (1972). Binary suprasegmental features and transformational word accentuation rules. *Language*, 48, 819-838.
- Vatikiotis-Bateson, E., & Kelso, J. A. S. (1993). Rhythm type and articulatory dynamics in English, French and Japanese. *Journal of Phonetics*, 21(3), 231-265.
- Verluyten, P. (1982). *Recherches sur la prosodie et la métrique du français*. Unpublished doctoral thesis, Universitaire Instelling Antwerp, Antwerp.
- Villing, R., Ward, T., & Timoney, J. (2003). *P-Centre Extraction from Speech: The need for a more reliable measure*. Paper presented at ISSC 2003, Limerick, Ireland, 1st-2nd July 2003.
- Voillat, F. (1971). *Aspects du français régional actuel*. Actes du colloque de dialectologie francoprovençale, Neuchâtel 1969.
- Volín, J., & Pollák, P. (2009). *The Dynamic Dimension of the Global Speech-Rhythm Attributes*. Paper presented at Interspeech 2009, Brighton, UK.

- Vrooman, J., Tuomainen, J., & de Gelder, B. (1998). The roles of word stress and vowel harmony in speech segmentation. *Journal of Memory and Language*, 38, 133-149.
- Vrooman, J., van Zon, M., & de Gelder, B. (1996). Cues to speech segmentation: Evidence from juncture misperceptions and word spotting. *Memory & Cognition*, 24, 744-755.
- Wagner, P. (2007). *Visualizing Levels of Rhythmic Organization*. Paper presented at the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany, 6th-10th August 2007.
- Wagner, P., & Dellwo, V. (2004). *Introducing YARD (Yet Another Rhythm Determination) and Re-introducing Isochrony to Rhythm Research*. Paper presented at Speech Prosody 2004, Nara, Japan, 23rd-26th March 2004.
- Walker, D. C. (2001). *French Sound Structure*. Alberta: University of Calgary Press.
- Wallin, J. E. W. (1901). Researches on the rhythm of speech. *Studies from the Yale Psychological Laboratory*, 9, 1-142.
- Wallin, J. E. W. (1911). Experimental studies of rhythm and time. *Psychological Review*, 18, 100-119.
- Walter, H. (1982). *Enquête phonologique et variétés régionales du français*. Paris: Presses universitaires de France.
- Walter, H. (1988). *Le français dans tous les sens*. Paris: R. Laffont.
- Wang, W. S.-Y., Lehiste, I., Chuang, C.-K., & Darnovsky, N. (1976). Perception of vowel duration. *Journal of the Acoustical Society of America*, 60(S1), S92.
- Warner, N., & Arai, T. (2001). Japanese mora-timing: a review. *Phonetica*, 58, 1-25.
- Welby, P. (2006). French intonational structure: Evidence from tonal alignment. *Journal of Phonetics*, 34, 343-371.
- Wenk, B. J. (1986). Cross-linguistic influence in second language phonology: Speech rhythm. In E. Kellerman & M. Sharwood-Smith (Eds.), *Cross-linguistic Influence in Second Language Acquisition* (pp. 120-133). New York and Oxford: Pergamon Institute of English.
- Wenk, B. J., & Wioland, F. (1982). Is French really syllable-timed? *Journal of Phonetics*, 10, 193-216.
- Werlen, I. (2000). *Der zweisprachige Kanton Bern*. Bern: Haupt.
- Werlen, I. (2002). *Vier Sprachen – eine Nation?* Paper presented at ‘Sprachen und Kulturen: viersprachig, mehrsprachig, vielsprachig. Langues et cultures: la Suisse, un pays où l’on parle quatre langues....et plus’, Biel-Bienne, Switzerland, 14th November 2002.
- White, L., & Mattys, S. (2007a). Calibrating rhythm: First language and second language studies. *Journal of Phonetics*, 35(4), 501-522.

- White, L., & Mattys, S. (2007b). Rhythmic typology and variation in first and second languages. In P. Prieto, J. Mascaró & M.-J. Solé (Eds.), *Segmental and Prosodic issues in Romance Phonology* (pp. 237-257). Amsterdam: John Benjamins.
- White, L., Mattys, S., Series, L., & Gage, S. (2007). *Rhythm metrics predict rhythmic discrimination*. Paper presented at the 16th International Congress of Phonetic Sciences, Saarbrücken, Germany, 6th-10th August 2007.
- Whitehall, H., & Hill, A. (1958). A report on the Language-Literature Seminar. In H. B. Allen (Ed.), *Readings in Applied English Linguistics* (pp. 393-397). New York.
- Whitworth, N. (2002). Speech rhythm production in three German-English bilingual families. *Leeds Working Papers in Linguistics and Phonetics*, 9, 175-205.
- Wickelgren, W. A. (1967). Rehearsal grouping and hierarchical organization of serial position cues in short-term memory. *Quarterly Journal of Experimental Psychology*, 19, 97-102.
- Widget, L., White, L., Schuppler, B., Grenon, I., Rauch, O., & Mattys, S. L. (2010). How stable are acoustic metrics of contrastive speech rhythm? *Journal of the Acoustical Society of America*, 127, 1559-1569.
- Wiese, R. (2000). *The Phonology of German*. Oxford: Oxford University Press.
- Woehrling, C., & Boula de Mareüil, P. (2006). Identification d'accents régionaux en français: perception et categorization. *Bulletin PFC: Phonologie du Français Contemporain – Usages, Variétés et Structure*, 6, 89-102.
- Woehrling, C., Boula de Mareüil, P., Adda-Decker, M., & Lamel, L. (2008). *A corpus-based prosodic study of Alsatian, Belgian and Swiss French*. Paper presented at Interspeech 2008 Brisbane, Australia, 22nd-26th September 2008.
- Woodrow, H. H. (1911). The role of pitch in rhythm. *Psychological Review*, 18, 54-77.
- Woodrow, H. H. (1909). *A quantitative study of rhythm; the effect of variations in intensity, rate and duration*. New York: The Science Press.
- Yu, A. C. L. (to appear). Tonal effects on perceived vowel duration. In C. Fougeron, B. Kühnert, M. D'Imperio & N. Vallée (Eds.), *Laboratory Phonology 10*. Berlin: Mouton de Gruyter.
- Zellner Keller, B. (2002). *Revisiting the Status of Speech Rhythm*. Paper presented at Speech Prosody 2002, Aix-en-Provence, France, 11th-13th April 2002.
- Zimmermann, G. (1998). Die „singende“ Sprechmelodie im Deutschen. *Zeitschrift für Germanistische Linguistik*, 26, 1-16.