Unsupervised intralingual and cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction

Matthew Gibson and William Byrne

***Abstract*—Hidden Markov model (HMM)-based speech synthesis systems possess several advantages over concatenative synthesis systems. One such advantage is the relative ease with which HMM-based systems are adapted to speakers not present in the training dataset. Speaker adaptation methods used in the field of HMM-based automatic speech recognition (ASR) are adopted for this task. In the case of unsupervised speaker adaptation, previous work has used a supplementary set of acoustic models to estimate the transcription of the adaptation data. This paper firstly presents an approach to the unsupervised speaker adaptation task for HMM-based speech synthesis models which avoids the need for such supplementary acoustic models. This is achieved by defining a mapping between HMM-based synthesis models and ASR-style models, via a two-pass decision tree construction process. Secondly, it is shown that this mapping also enables unsupervised adaptation of HMM-based speech synthesis models without the need to perform linguistic analysis of the estimated transcription of the adaptation data. Thirdly, this paper demonstrates how this technique lends itself to the task of unsupervised *cross-lingual* adaptation of HMM-based speech synthesis models, and explains the advantages of such an approach. Finally, listener evaluations reveal that the proposed unsupervised adaptation methods deliver performance approaching that of supervised adaptation.**

***Index Terms*—HMM-based speech synthesis, unsupervised speaker adaptation, cross-lingual.**

I. Introduction

HIDDEN Markov model-based systems have delivered synthetic speech of similar quality to that of concatenative (or unit selection) synthesis systems [1]. Additionally, HMM-based systems possess several advantages over unit selection systems. These advantages include simple ways to interpolate between speakers and synthesise speech of varying styles or emotions [2; 3]. Perhaps the most significant advantage is the ability to adapt to new speakers using a relatively small amount of adaptation data [4].

Most research into speaker adaptation for HMM-based speech synthesis (or text-to-speech, TTS) has focussed upon the supervised scenario, where transcribed adaptation data is available. More recent work has tackled the challenge of adaptation of HMM-based synthesis models using unlabelled adaptation data [5]. As will be explained in due course, unsupervised adaptation of HMM-based synthesis models is problematic for two reasons. Firstly, the modelling of suprasegmental contextual information renders the synthesis models unsuitable for ASR purposes. Therefore a supplementary set of triphone acoustic models are typically used to estimate a transcription of the unlabelled adaptation data [5]. Secondly, linguistic analysis is required to transform word-level transcriptions into transcriptions containing suprasegmental contextual information. In the case of unsupervised adaptation, it is feasible that such linguistic analysis exacerbates errors present in the estimated word-level transcription.

This paper presents an alternative to the unsupervised adaptation approach described in [5]. In [5], adaptation transforms estimated using triphone acoustic models are applied to the more detailed acoustic models typically used in HMM-based synthesis. While this technique avoids the need for linguistic analysis of the estimated transcription of the adaptation data, a separately-estimated triphone acoustic model set is still required.

In this paper, a two-stage decision tree construction method is introduced, which enables a single set of acoustic model components to be used for both ASR and TTS. This method is then used to circumvent the need for supplementary ASR acoustic models and linguistic analysis of estimated transcriptions during unsupervised adaptation. The application of the two-stage decision tree construction method is then extended to the task of unsupervised cross-lingual speaker adaptation.

Cross-lingual (or interlingual) speaker adaptation is defined as the adaptation of acoustic models associated with one language, the *target language*, using adaptation data uttered in a different language, the *source language*.

A large amount of research has been performed on the cross-lingual adaptation task for ASR acoustic models. The task typically arises in cases where a relatively small amount of data is available to train an ASR acoustic model in a particular target language. Bootstrapping the target language acoustic models ([6]) based upon an explicit mapping from source to target language phonemes has been explored, as well as interpolation of the source and target language acoustic models (also [6]). Later work ([7]) has successfully applied the maximum a-posteriori (MAP) adaptation method to the cross-lingual adaptation task, demonstrating the usefulness of the prior knowledge contained within the source language.

Recent work [8; 9] has addressed the task of supervised cross-lingual adaptation for HMM-based speech synthesis. This work used TTS models of both source and target languages, and defined a phoneme or state-level mapping between the source and target language acoustic models. This mapping was then deployed to translate the source language transcription of the adaptation data to a target language phoneme or state sequence. The target language TTS models were subsequently adapted using the source language acoustic data and the corresponding mapped target language phoneme or state sequence.

Techniques similar to those described above rely upon the availability of both source and target language TTS models, and the mapping mechanism between these models must be established prior to adaptation. An alternative approach based upon the two-stage decision tree construction technique is proposed in this paper. As will be explained later, this alternative approach is appealing because it requires no knowledge of the source language acoustic model (or even the source language) or its relationship to the target language acoustic model.

This paper evaluates the proposed unsupervised adaptation schemes in both a standard adaptation scenario and a speaker adaptive training (SAT) framework. The performance of these techniques is compared with standard approaches to supervised and unsupervised speaker adaptation of HMM-based synthesis models in both the *intralingual* (within-language) and cross-lingual scenarios. In the cross-lingual case, parallel translated adaptation datasets recorded by the same speaker are used to compare the performance of intralingual and cross-lingual adaptation in a controlled manner. Listener evaluations reveal that the proposed unsupervised adaptation techniques deliver performance approaching that of supervised intralingual adaptation.

The paper is structured as follows. Section II provides a brief introduction to HMM-based speech synthesis models and explains why unsupervised adaptation of such models is problematic. Section III explains the two-pass decision tree construction technique, and how this enables unsupervised adaptation of HMM-based synthesis models. Sections IV and V respectively introduce the unsupervised intralingual and cross-lingual approaches used in this work. Section VI discusses the use of SAT in the context of HMM-based speech synthesis. The proposed approaches to intralingual and cross-lingual speaker adaptation are evaluated in Sections VII and VIII respectively. Lastly, Section IX summarises the contributions of this work and highlights areas of future research.

II. Unsupervised adaptation and HMM-based speech synthesis

In the domain of ASR, unsupervised adaptation is usually conducted by firstly estimating a transcription of the adaptation data using a speech recogniser. This speech recogniser usually deploys the same models which are subsequently adapted.

In the domain of HMM-based synthesis, use of the same unsupervised adaptation framework is problematic because the acoustic models typically used in HMM-based speech synthesis are not easily integrated into the ASR search procedure. This, in turn, is because the context-dependent acoustic models used in HMM-based speech synthesis [10] represent suprasegmental information (e.g. syllabic stress, total number of syllables in utterance) in addition to segmental information (e.g. context-dependent phoneme label). These models are henceforth referred to as *full context* models. Although theoretically possible to recognise unlabelled data using full context models, this requires information which relates to complete hypotheses (e.g. the total number of words in an utterance) when constructing a recognition network. When using e.g. triphone acoustic models, such information may be ignored to simplify the recognition network and to facilitate dynamic network construction. The presence of suprasegmental contextual information in full context models therefore adds a prohibitive amount of complexity to the construction of recognition networks.

A simple solution to this problem is to use a separately-estimated ASR-compliant acoustic model to obtain a transcription of the adaptation data, followed by adaptation of the TTS model using this transcription [5]. However this solution involves estimation of a separate ASR model, and such model estimation is often a lengthy procedure. Further, use of different models during the recognition and adaptation stages precludes the use of efficient online adaptation strategies [11]. For these reasons, alternative techniques which enable TTS models to be deployed for ASR have been explored [12]. The two-pass decision tree construction technique [13] is one such technique, as will be explained in the following section.

III. Two-pass decision tree construction

As is the case for ASR acoustic modelling, decision tree clustering of the full contexts is used to enable robust estimation of the model parameters. The minimum description length (MDL) criterion [14] is used when constructing the decision tree, which in turn uses questions pertaining to both segmental and suprasegmental context. By performing this decision tree construction in two stages, where the initial stage uses questions relating to triphone contextual information, and the second stage uses questions relating to all contextual information, a well-defined mapping between full context models and triphone models may be established. This constrained decision tree construction process is illustrated in Figure 1.

The first stage, indicated as Pass 1 in Figure 1, uses only questions relating to left, right and central phonemes to construct a phonetic decision tree. This decision tree is used to generate a set of tied triphone contexts, which are easily integrated into the ASR search. No state output distributions are estimated at this stage. Pass 2 extends the decision tree constructed in Pass 1 by introducing additional questions relating to suprasegmental information. The output of Pass 2 is an extended decision tree which defines a set of tied full contexts. The MDL criterion is used for both Pass 1 and Pass 2.

After this two-pass decision tree construction, single component Gaussian state output distributions are estimated to model the tied full contexts associated with each leaf node of the extended decision tree. These models are then used for speech synthesis.

A mapping from the single component full context models to multiple component triphone models is defined as follows. Each set of Gaussian components associated with the same ‘triphone ancestor’ are grouped to form a multiple component mixture distribution to model the triphone context defined by the ‘triphone ancestor’. The derived triphone models are illustrated at the bottom of Figure 1. The mixture weight of each component is calculated from the occupancies associated with the Pass 2 leaf node contexts.

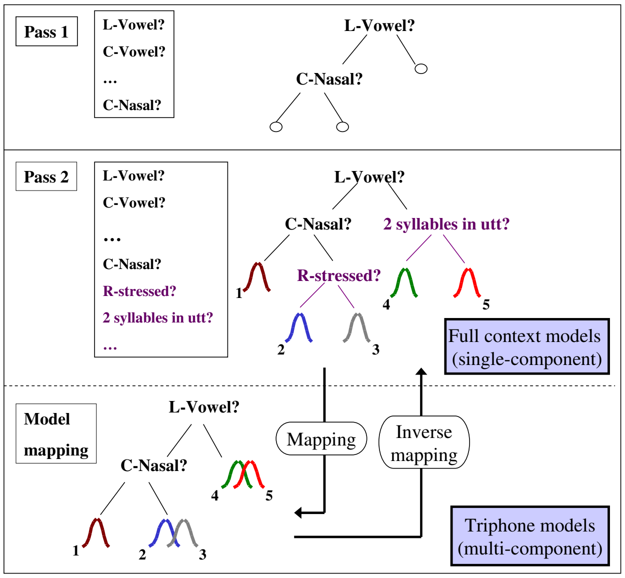

Fig. 1. Two-pass decision tree construction. Mapping functions permit sharing of components between full context models for TTS and triphone models for ASR.

The inverse mapping from triphone models to full context models is obtained by associating each Gaussian component with its original full context. This is achieved by assigning a unique full context identifier to each component as illustrated in Figure 1.

Mapping full context models to triphone models enables ASR compatible acoustic models to be derived from TTS acoustic models, thus avoiding the need for a separately-estimated ASR model. Sections IV and V explain how these mappings between full context and triphone models can be exploited to perform unsupervised intralingual and cross-lingual adaptation of full context models.

IV. Unsupervised Intralingual Adaptation

As illustrated in Figure 2, triphone models derived from estimated full context models (as described in Section III) are used to transcribe unlabelled adaptation data. One question remains, however. How is ASR output, e.g. a word, phoneme or triphone sequence, used to adapt full context models? One method, labelled as 'full adaptation' in Figure 2, firstly performs linguistic analysis of the estimated word-level transcription to produce an estimated full context labelling of the adaptation data. The full context models are then adapted directly using this labelling.

By defining an inverse mapping between full context and triphone models, the two-pass decision tree construction method introduces an alternative to the 'full adaptation' technique. As illustrated in Figure 2, the estimated triphone transcription may be used to adapt the triphone models. The adapted triphone models are then subsequently mapped back to full context models using the inverse mapping. This is labelled as 'triphone adaptation' in Figure 2.

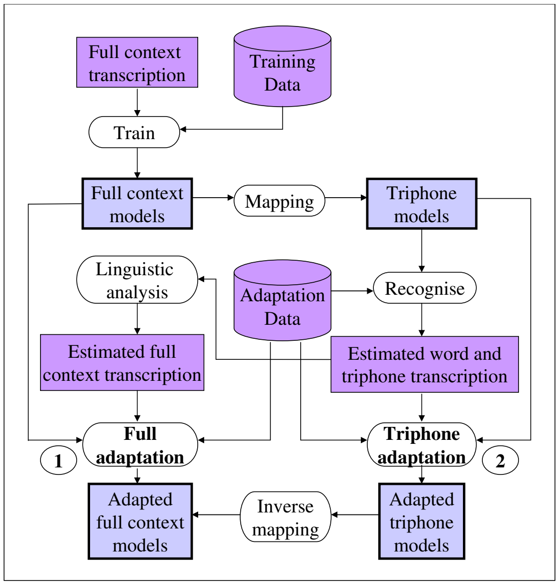

Fig. 2. Unsupervised adaptation of full context models via (1) full adaptation or (2) triphone adaptation.

Once word and triphone-level transcriptions of the adaptation data are available, the full context models may be adapted in these two different ways. Note that linguistic analysis may exacerbate errors present in the estimated word-level transcription. It is therefore feasible that the triphone adaptation technique is more robust than full context adaptation in the unsupervised case. This hypothesis is tested in the experiments of Section VII.

V. Unsupervised Cross-Lingual Adaptation

Consider now the task of unsupervised cross-lingual speaker adaptation, as defined in Section I, in the case of full context acoustic models. To transcribe the adaptation data one could deploy an ASR system tailored to the source language i.e. a source language lexicon, as well as source language acoustic and language models. This estimated transcription may then be subsequently mapped to the target language. This mapping may be defined at the phone level [8] or the state level [9]. The mapped transcription may then be used to adapt the target language full context models.

The above approach deploys a large amount of source language specific knowledge, as well as knowledge of the relationship between source and target languages. Acquisition of such knowledge typically depends upon a large amount of transcribed acoustic data from the source language. Such a database is certainly not available for all languages, and is expensive to obtain. Further, if the source language is unknown, clearly the approach described above cannot be applied. For these reasons, an alternative method is explored in this work.

The cross-lingual adaptation technique used in this work treats the source language adaptation data as if it were uttered in the target language. Target language acoustic models and a phoneme loop grammar are used to recognise the adaptation data, thus mapping it onto a phoneme sequence in the target language. Subsequently, the estimated triphone sequence is used as the reference sequence, and the triphone adaptation method of Figure 2 is used. This process is almost identical to the triphone adaptation approach to unsupervised intralingual adaptation. The sole difference is that, in order to avoid language specific constraints, no dictionary or language model is used during recognition. This method was first introduced and evaluated in [15].

The approach described in the previous paragraph uses no source language ASR or TTS system. Further, no previously learned mapping between source and target language acoustic models is necessary. Indeed, no source language knowledge whatsoever is used, so the technique may be applied even when the source language is unknown.

By comparing the performance of unsupervised intralingual and cross-lingual adaptation, the impact of source language knowledge may be measured. This comparison is reported in Section VIII.

VI. Speaker adaptive training

Speaker adaptive training (SAT, [16]) attempts to decouple inter-speaker and intra-speaker variance when estimating a speaker independent (SI) acoustic model. The SAT framework simultaneously estimates sets of speaker dependent transforms of the acoustic models (one set of transforms for each speaker in the training set) and a speaker independent 'canonical' model. The transforms are designed to capture much of the inter-speaker acoustic variance and consequently the canonical model displays less variance than a standard SI system.

Both SAT-estimated and standard SI full context models are used in the experimental work of this paper. Figure 3 illustrates the procedure used to estimate these models. SAT-estimated monophone models are estimated, then cloned to full context models, which are SAT-estimated using one global transform per state/stream combination per speaker. The statistics from these untied full context models are then used to cluster the full context models. Subsequent to full-context clustering, tied models are re-estimated to create both SAT-estimated and standard SI tied full context models.

There is evidence [17] that SAT-estimated models are superior to standard SI-estimated models for HMM-based speech synthesis. The evaluation of Section VIII revisits this comparison to determine if the same conclusions hold in the case of models generated using two-pass decision tree construction. The performance of SAT-estimated and standard SI models is compared both prior to and after adaptation.

VII. Evaluation: intralingual speaker adaptation

The evaluation described in this section is designed to address the following questions regarding unsupervised intralingual speaker adaptation of HMM-based synthesis models.

1) Does the constrained two-pass decision tree construction process affect the naturalness of the resulting speech?
2) How does the proposed approach to unsupervised intralingual adaptation compare with supervised intralingual adaptation?
3) How does the performance of triphone adaptation (as described in Section IV) compare with that of full adaptation?

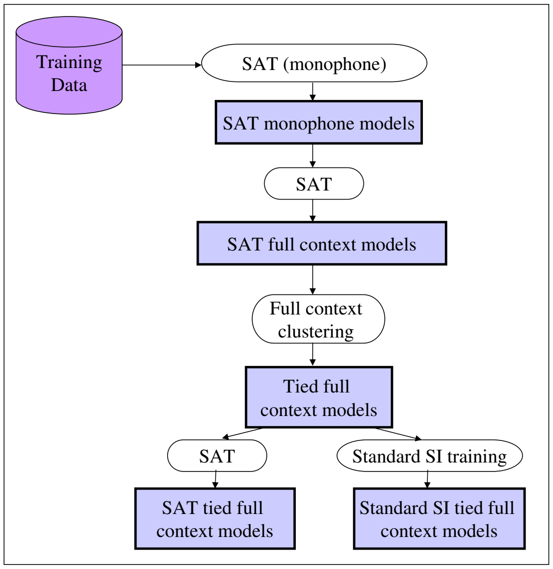

Fig. 3. Estimation of speaker independent (SI) full context models using speaker adaptive training (SAT) and standard model estimation.

A. Background information

The synthesis models used in this evaluation deploy the following acoustic features: STRAIGHT-analysed Mel-cepstral coefficients [18] (40 dimensions), fundamental frequency ($F0$), and measurements which quantify the aperiodicity of the speech (5 dimensions). The first and second order temporal derivatives of all of these coefficients are appended to yield a feature vector of dimension 138. The feature vector is split into five streams: cepstral coefficients, aperiodicity measures, $F0$, first derivative of $F0$, and second derivative of $F0$. Multi-space probability distributions are used to model observations of varying dimension, namely the $F0$ observation [19]. Explicit duration models (hidden semi-Markov models, [20]) are integrated to improve the quality of synthesised speech. One decision tree per state and stream combination (where all three $F0$ streams are combined for the purposes of clustering) is used, with an additional decision tree to cluster contexts of the duration model. A speech utterance is generated from full context models via feature sequence generation with global variance consideration [21; 22]. Synthesis of the waveform from the feature sequence is performed by the STRAIGHT vocoder [18].

B. Systems

To address the questions posed at the start of this section, the systems detailed in Table I are evaluated. Standard SI full context models are estimated using the Wall Street Journal (WSJ) SI84 training dataset ($3,586$ male and $3,552$ female utterances, 7136 utterances) and maximum likelihood estimation. Note that such databases have proven useful for HMM-based speech synthesis ([23]).

System	Clustering	Adaptation method	Supervised?
A	Standard	-	-
B	Two-pass	-	-
C	Two-pass	Full	Y
D	Two-pass	Full	N
E	Two-pass	Triphone	Y
F	Two-pass	Triphone	N
G	-	-	-

TABLE I
EVALUATED SYSTEMS (INTRALINGUAL ONLY).

Average voice models corresponding to standard, unconstrained decision tree construction (system A of Table I) are estimated for comparison with those corresponding to two-pass decision tree construction (system B). Note that only Mel-cepstral, $F0$, and aperiodicity models are adapted in this work, so only those models are clustered using the two-pass decision tree construction method. Duration models are clustered using standard clustering methods and are identical in systems A and B.

Adapted systems are derived from System B using either the triphone or full adaptation method described in Section IV. Constrained maximum likelihood linear regression (CMLLR, [24]) adaptation is used, and the adaptation data corresponds to spoke 4 of the 1993 ARPA evaluation (40 utterances for speaker 440M). The adaptation techniques are evaluated in the supervised and unsupervised cases, resulting in four adapted model sets corresponding to systems C through F in Table I.

System G corresponds to vocoded natural speech, analysed and resynthesised using the STRAIGHT technique [18]. This system is included in the evaluation to establish an upper bound on the performance of the synthesised speech.

In the case of unsupervised adaptation, triphone models derived from the estimated full context average voice models are used for the recognition step, in conjunction with the closed vocabulary 20k bigram language model provided with the WSJ0 corpus. A set of state transition probabilities are estimated from the SI84 dataset for use with the triphone models during recognition. A phoneme error rate of 47.1% is observed for the unsupervised transcriptions.

C. Analysis of two-pass decision tree construction

Table II displays the number of leaf nodes created using different decision tree construction methods, and for the different streams. In all cases, the number of leaf nodes generated after two-pass decision tree construction exceeds that of standard tree construction. This demonstrates that constraining the tree structure to satisfy the requirements of the two-pass construction method, defined in Section III, leads to less compact trees.

	Mel-cepstral	$F0$	aperiodicity
Pass 1	2208	6756	1644
Pass 2	2889	34849	2639
Standard	2621	30581	2160

TABLE II
NUMBER OF LEAF NODES CREATED USING DIFFERENT DECISION TREE CONSTRUCTION METHODS.

D. Evaluation details

Two different evaluation methods were used to measure the performance of the two-pass intralingual adaptation technique: an opinion score evaluation, described in Section VII-D1, and a paired comparison of several pairs of systems, described in Section VII-D2. The opinion score evaluation provides a performance measure and overall ranking of the systems studied, while the paired comparison more effectively discovers significant differences between system pairs.

1) Opinion score evaluation: The seven systems (A through G) were evaluated by listening to synthesised utterances via a web browser interface closely resembling that used in the Blizzard Challenge 2007. The evaluation comprised two sections. In the first section, listeners judged the naturalness of an initial set of synthesised utterances. In the second section, listeners judged the similarity of a second set of synthesised utterances to a target speaker's (speaker 440M) speech. Four of the target speaker's natural utterances were available for comparison. No utterances from the initial set were present in the second set. Each synthetic utterance was judged using a five point Likert-type psychometric response scale [25], where '5' is the most favourable response and '1' is the least favourable.

Twenty two native English speakers conducted the evaluation. A Latin square experimental design was used to define the order in which systems were judged (a different square for each section of the evaluation). Each listener was assigned a row of each Latin square, and judged seven different utterances per section, each synthesised by a different system. The synthesised utterances are a subset of the 1992 ARPA speaker independent read 5k test dataset with no verbal punctuation.

Throughout this paper, significant differences between systems evaluated using the opinion score evaluation are detected using a pairwise Wilcoxon signed rank test which is Bonferroni-corrected for multiple comparisons [26]. A difference is deemed significant if this test discovers significance at the 95% confidence level.

2) Paired comparison evaluation: Three pairs of systems are selected and a preference test conducted in order to address the questions stated at the start of this section. Each judge was presented with pairs of synthesised utterances, one generated from each system in the comparison. For each pair, the judge was forced to select his preferred system, according to either naturalness or similarity to a target speaker. In the case of similarity, four of the target speaker's natural utterances were available to inform the judgement. The synthesised utterances are a subset of the 1992 ARPA speaker independent read 5k test dataset with no verbal punctuation.

The following pairs of systems were compared. Unadapted standard (system A) and unadapted two-pass (system B) were compared in terms of naturalness. Supervised triphone-adapted (system E) and unsupervised triphone-adapted (system F) were compared in terms of target speaker similarity. Lastly, supervised full-adapted (system C) and unsupervised triphone-adapted (system F) were also compared in terms of target speaker similarity. Thirty-four pairs of utterances were presented in each comparison. Ten native English speakers conducted the evaluation.

Throughout this paper, significant differences between systems evaluated using the paired comparison method are detected using Pearson's chi-square test to approximate the binomial test. A difference is deemed significant if this test discovers significance at the 95% confidence level.

E. Results

1) Opinion scores: Figure 4 summarises listener judgements of 'naturalness' and 'similarity to target speaker' using boxplots. Table III displays the corresponding mean opinion scores (MOS) for naturalness and similarity for each system.

System	MOS naturalness	MOS similarity
A	2.0	1.0
B	1.8	1.0
C	2.1	3.3
D	1.9	2.8
E	2.1	2.9
F	2.0	2.9
G	3.8	4.9

TABLE III
MEAN OPINION SCORES FOR NATURALNESS AND SIMILARITY TO TARGET SPEAKER.

Significant differences are observed between vocoded natural speech (System G) and all other systems for both naturalness and similarity to the target speaker. Listeners clearly assign lower scores to synthetic speech.

With regard to target speaker similarity, significant differences are observed between the unadapted models (system B) and all adapted systems (C through F). No significant difference is observed between any pair of adapted systems.

With regard to naturalness, no significant differences are observed between any of the synthetic speech systems (A through F).

2) Paired comparisons: Table IV displays the frequency with which each system was preferred in the paired comparisons evaluation described in Section VII-D2. A significant difference in naturalness is detected between the standard unadapted system (system A) and the unadapted two-pass system (system B). A significant difference in target speaker similarity is also detected between the supervised full-adapted system (system C) and the unsupervised triphone-adapted system (system F). No significant difference is detected between supervised and unsupervised triphone-adapted systems (systems E and F respectively).

System (% of times selected)		Significant difference?
A (57.1%)	B (42.9%)	Y
E (55.0%)	F (45.0%)	N
C (72.6%)	F (27.4%)	Y

TABLE IV
PAIRED COMPARISONS FOR INTRALINGUAL ADAPTATION METHODS. EACH ROW CORRESPONDS TO A SINGLE COMPARISON.

F. Discussion

The questions phrased at the start of this section will now be addressed in turn.

1) Does the constrained two-pass decision tree construction process affect the naturalness of the resulting speech?: A small but significant decrease in naturalness is observed between system A (standard decision tree construction) and system B (two-pass decision tree construction). So constraining decision tree construction using the two-pass technique has compromised the naturalness of the resulting synthetic speech. However this is only a small loss in naturalness, as shown in Table III.

2) How does the proposed approach to unsupervised intralingual adaptation compare with supervised intralingual adaptation?: In the case of full adaptation, a reasonably large reduction in target speaker similarity MOS from 3.3 (system C) to 2.8 (system D) is observed when using unsupervised adaptation. In the case of triphone adaptation, supervised (system E) and unsupervised (system F) methods deliver the same target speaker similarity MOS of 2.9, and no significant difference is found between these systems in a paired comparison test.

To summarise, in the case of full adaptation there is evidence to suggest that supervised adaptation delivers superior performance to the unsupervised case. A significant difference is found between unsupervised triphone adaptation (system F) and supervised full adaptation (system C). However, in general, these results demonstrate that unsupervised adaptation of TTS models achieves performance approaching that of supervised adaptation. This is achieved without use of supplementary acoustic models or any source-language training material.

3) How does the performance of triphone adaptation (as described in Section IV) compare with that of full adaptation?: In the supervised case, full adaptation (system C) delivers a superior target speaker similarity MOS to that of triphone adaptation (system E).

In the unsupervised case, the opposite is true. Triphone adaptation (system F) delivers a superior target speaker similarity MOS to that of full adaptation (system D).

These results suggest that there is a relationship between the optimal choice of adaptation technique (full or triphone) and the quality of the estimated transcription of the adaptation data. In the unsupervised case here, it has been demonstrated that linguistic analysis of the adaptation data may be bypassed by using triphone adaptation without adversely affecting the unsupervised adaptation procedure.

VIII. EVALUATION: CROSS-LINGUAL SPEAKER ADAPTATION

The evaluation described in this section is designed to address the following questions in the context of unsupervised

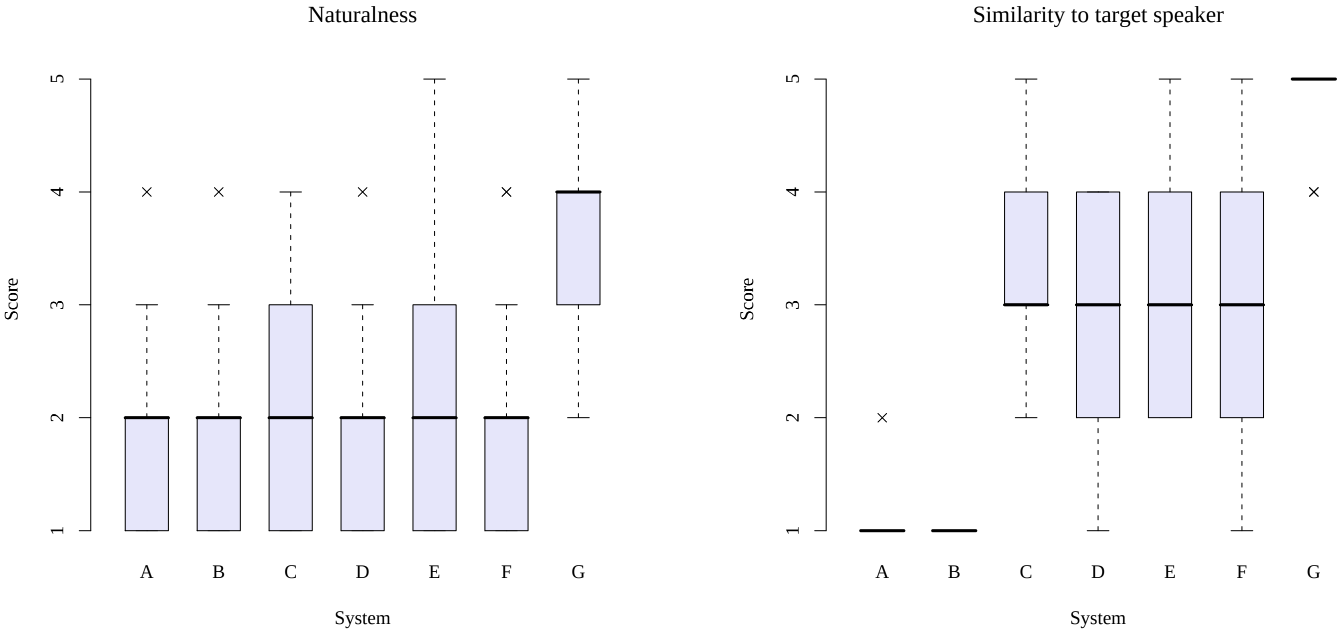

Fig. 4. Boxplots of listener opinion scores for naturalness and similarity to target speaker.

cross-lingual speaker adaptation of HMM-based synthesis models.

1) How does the proposed approach to unsupervised cross-lingual adaptation compare with unsupervised intralingual adaptation?
2) Can knowledge of the source language improve the quality of unsupervised cross-lingual adaptation?
3) How well does the unsupervised cross-lingual adaptation technique generalise across speakers and languages?

Additionally, the following questions, related to the SAT framework, will be addressed in the context of supervised intralingual speaker adaptation.

4) Does SAT estimation improve the quality of the unadapted and adapted models?
5) Does the constrained two-pass decision tree construction process affect the quality of the resulting SAT-estimated models?

The background information detailed in Section VII-A remains relevant for this evaluation.

A. Adaptation datasets

The speech of five competent, but not native, male speakers of English is used as adaptation data. In the case of the European languages considered (French, Italian, Dutch and Finnish), this speech corresponds to utterances selected from the Europarl corpus of parallel text of European parliament proceedings [27]. In the case of Mandarin, the speech corresponds to a subset of the NIST 2008 Chinese-English MT evaluation parallel texts. Each speaker provided speech in his native language as well as the parallel translated speech in English. The semantic content of the data is therefore identical in both languages. Additionally, the parallel speech data was recorded in the same acoustic environment. The use of adaptation datasets corresponding to the same speaker, semantics, and acoustic conditions enables a controlled comparison of intralingual and cross-lingual speaker adaptation.

Statistics relating to the adaptation datasets for each speaker/language pair are provided in Table V. The datasets were designed to correspond to approximately the same total number of English words.

Speaker native language	Adaptation data language	# utterances	# minutes	# words
Mandarin	Mandarin English	84	8.2 10.0	- 1221
French	French English	113	10.8 9.4	1353 1227
Italian	Italian English	119	11.9 12.8	1312 1226
Dutch	Dutch English	113	10.2 10.1	1123 1222
Finnish	Finnish English	71	8.3 9.6	839 1224

TABLE V
PARALLEL ADAPTATION DATASETS.

B. Systems

The questions highlighted at the start of this section inform the choice of systems to be evaluated. In total, eight systems, detailed in Table VI are evaluated.

English male average voice full context models are generated using the male-only subset of the Wall Street Journal (WSJ) SI84 training dataset ($3,586$ utterances). Using two-pass decision tree construction, both SAT (system I) and standard SI (system J) models are estimated as discussed in Section VI. For comparison, a SAT-estimated model is estimated using standard decision tree construction (system H). Again, note that only Mel-cepstral, $F0$, and aperiodicity models are adapted in this work, so only those models are clustered using the two-pass decision tree construction method.

Three SAT-adapted models (systems K through M) are derived from the SAT models (system I) using CMLLR adaptation. System K is the result of applying unsupervised cross-lingual adaptation as described in Section V, and using the native speech adaptation datasets described in Section VIII-A. System L is the result of unsupervised intralingual adaptation (triphone adaptation) as described in Section IV, and using the English speech adaptation datasets described in Section VIII-A. System M is identical to system L with the exception that the correct transcription is used during adaptation.

One standard adapted system (system N) is derived from system J, again using CMLLR. System N differs from system M only in that it is adapted from the standard SI models (system J). System O corresponds to vocoded natural speech.

System	Clustering	SI model estimation method	Source language	Supervised?
H	Standard	SAT	-	-
I	Two-pass	SAT	-	-
J	Two-pass	Standard SI	-	-
K	Two-pass	SAT	Native	N
L	Two-pass	SAT	English	N
M	Two-pass	SAT	English	Y
N	Two-pass	Standard SI	English	Y
O	-	-	-	-

TABLE VI
EVALUATED SYSTEMS (CROSS-LINGUAL AND INTRALINGUAL).

C. Evaluation details

As in Section VII-D, two different evaluation methods were used to measure the performance of the cross-lingual adaptation technique: an opinion score evaluation, described in Section VIII-C1, and a paired comparison of several pairs of systems, described in Section VIII-C2.

1) Opinion score evaluation: As described in Section VII-D, all systems were evaluated by rating synthesised utterances via a web browser interface using a five point psychometric response scale. These utterances are English sentences extracted from the Europarl corpus, and they are distinct from the adaptation utterances.

The evaluation comprised ten sections. In the first set of five sections (one per speaker), listeners judged the naturalness of an initial set of forty synthesised utterances. In the second set of five sections (again, one per speaker), listeners judged the similarity of a second set of forty synthesised utterances to the target speaker's speech. In each section, four of the target speaker's natural English utterances were available for comparison.

Twenty-four paid judges conducted the evaluation. Twenty-one were native English speakers and the remaining three had spent more than two years in an English-speaking country at the time of the evaluation. Different Latin squares were used for each section to define the order in which systems were judged. Each listener was assigned a row of each Latin square, and judged eight different utterances per section, each synthesised by a different system.

2) Paired comparison evaluation: Two pairs of systems are selected and a preference test similar to that described in Section VIII-C2 was conducted in order to address some of the questions stated at the start of this section. The following pairs of systems were compared: unsupervised intralingual adapted (system L) and unsupervised cross-lingual adapted (system K), and supervised intralingual adapted (system M) and unsupervised cross-lingual adapted (system K). Forty pairs of utterances were presented in each comparison. Ten native English speakers conducted the evaluation.

D. Results

1) Opinion scores: Figure 5 summarises listener judgements of target speaker similarity and naturalness using boxplots. Tables VII and VIII display, respectively, the average target speaker similarity and naturalness for each system in the column labelled 'av'.

Sys	MOS similarity					
	M	Fr	I	D	Fi	av
H	1.4	1.4	1.6	1.6	1.6	**1.5**
I	1.2	1.3	1.3	1.3	1.6	**1.3**
J	1.3	1.3	1.4	1.5	1.8	**1.4**
K	1.5	1.3	1.8	1.8	1.8	**1.7**
L	1.9	1.5	1.7	2.0	1.8	**1.8**
M	2.0	1.9	2.0	1.8	2.0	**1.9**
N	1.9	1.7	2.0	2.0	2.2	**2.0**
O	4.9	5.0	5.0	4.7	4.8	**4.9**

TABLE VII
MEAN OPINION SCORES FOR SIMILARITY TO TARGET SPEAKER, ANALYSED BY TARGET SPEAKER NATIVE LANGUAGE (M=MANDARIN, FR=FRENCH, I=ITALIAN, D=DUTCH, FI=FINNISH).

Sys	MOS naturalness					
	M	Fr	I	D	Fi	av
H	-	-	-	-	-	**2.9**
I	-	-	-	-	-	**2.8**
J	-	-	-	-	-	**2.8**
K	2.5	2.4	2.5	2.4	2.0	**2.4**
L	2.5	2.5	2.8	2.5	2.1	**2.5**
M	2.6	2.8	2.7	2.3	2.5	**2.6**
N	2.7	2.6	2.8	2.7	2.2	**2.6**
O	3.9	4.6	4.5	4.9	4.4	**4.4**

TABLE VIII
MEAN OPINION SCORES FOR NATURALNESS, ANALYSED BY TARGET SPEAKER NATIVE LANGUAGE (M=MANDARIN, FR=FRENCH, I=ITALIAN, D=DUTCH, FI=FINNISH).

Again, significant differences exist between vocoded natural speech (system O) and all other systems, both in terms of naturalness and target speaker similarity.

The average similarity to the target speaker given by SAT-adapted systems (K, L and M) are all significantly greater than

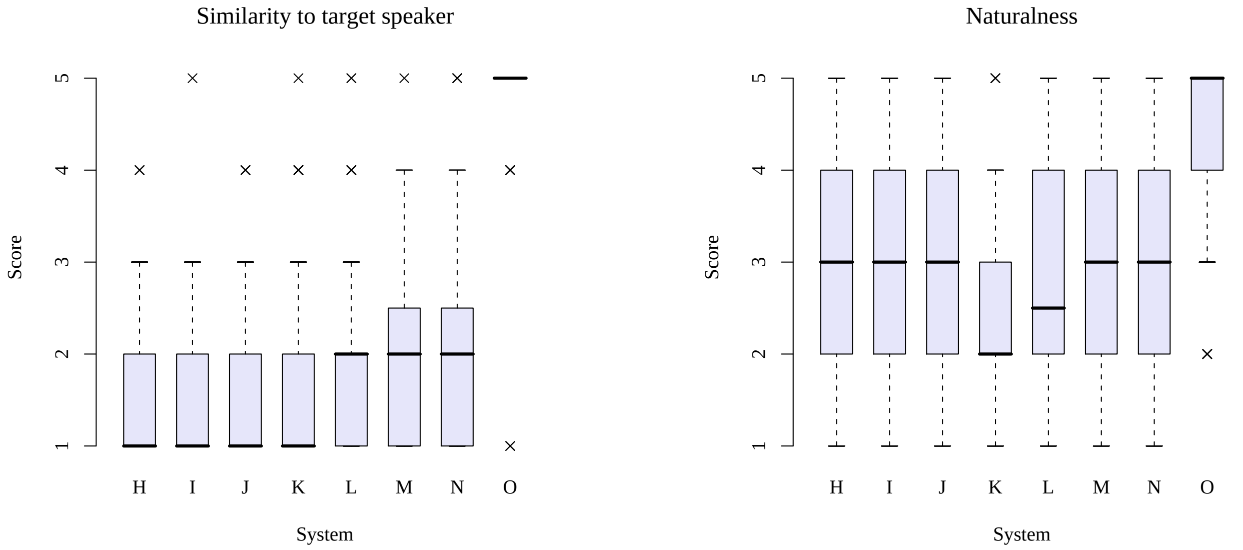

Fig. 5. Listener opinion scores for similarity to target speaker and naturalness.

that observed for the corresponding unadapted models (system I). No significant difference is observed between the similarity of any pair of SAT-adapted systems. The average similarity of the standard adapted models (system N) is significantly greater than that observed for the corresponding unadapted models (system J).

With regard to naturalness, although the adapted systems (K through N) display lower average scores, no significant difference is detected between the naturalness of any adapted system and its corresponding unadapted system.

2) Paired comparisons: Table IX displays the frequency with which each system was preferred in the paired comparisons evaluation described in Section VIII-C2. No significant difference in target speaker similarity is detected between the unsupervised intralingual adapted system (system L) and the unsupervised cross-lingual adapted system (system K). A significant difference in target speaker similarity is detected between the supervised intralingual adapted system (system M) and the unsupervised cross-lingual adapted system (system K).

System (% of times selected)		Significant difference?
L (53.2%)	K (46.8%)	N
M (56.7%)	K (43.3%)	Y

TABLE IX
PAIRED COMPARISONS FOR CROSS-LINGUAL ADAPTATION METHODS. EACH ROW CORRESPONDS TO A SINGLE COMPARISON.

E. Discussion

The results presented above are now discussed in relation to the questions specified at the start of this section.

1) How does the proposed approach to unsupervised cross-lingual adaptation compare with unsupervised intralingual adaptation?: The evidence presented above suggests that unsupervised intralingual adaptation yields superior performance to unsupervised cross-lingual adaptation: compare the average target speaker similarity in the case of system L (unsupervised intralingual adaptation, 1.8) to that of system K (unsupervised cross-lingual adaptation, 1.7). Note, however, that this evidence, and that of the paired comparison between these systems, is not sufficient to prove that any significant difference exists between the systems.

2) Can knowledge of the source language improve the quality of unsupervised cross-lingual adaptation?: System M (supervised intralingual adaptation) may be thought of as an unsupervised cross-lingual system with an ideal mapping from source language speech to target language phoneme sequence. As such, its performance, an average target speaker similarity of 1.9 and average naturalness of 2.6, provides a reasonable upper limit for the performance of cross-lingual speaker adaptation. This is superior performance to that observed for unsupervised cross-lingual adaptation (system K, similarity of 1.7 and naturalness of 2.4). Note also that a significant performance difference is found between these systems in a paired comparison, so it can be argued that use of source language knowledge may narrow the margin between these systems.

3) How well does the unsupervised cross-lingual adaptation technique generalise across speakers and languages?: Comparing the values for system I with those of system K in Table VII, it is observed that the unsupervised cross-lingual adaptation technique successfully increases target speaker similarity for all speakers/languages chosen in this study. While this demonstrates that the technique generalises well, note that this increase in similarity varies widely across speakers/languages. For example, in the case of Dutch, an increase of 0.5 is observed (from 1.3 to 1.8) while, in the case of French, an increase of less than 0.1 (from 1.29 to 1.33) is recorded.

This variance may be due to language-specific factors e.g. the extent of overlap between the phonetic inventory of the source language and that of English. Several other factors may contribute to this variance, however, e.g. varying volumes of adaptation data used (see Table V), differing phonetic content across adaptation datasets, or differences in recognition accuracy across speakers.

An alternative explanation should be kept in mind. The target speaker's characteristics may change when speaking his non-native English. When adapting using native language speech, such alterations are not observed, and so possibly not captured. This is a fundamental issue with cross-lingual speaker adaptation. However, this issue may be more pronounced for certain speakers or languages.

Further experimentation and analysis is required to explain the varying performance of the unsupervised cross-lingual speaker adaptation.

4) Does SAT estimation improve the quality of the unadapted and adapted models?: Both the unadapted SAT models (system I) and the standard SI models (system J) yield an average naturalness score of 2.8, as displayed in Table VIII. In the case of the adapted systems, both the standard adapted models (system N) and SAT-adapted models (system M) yield an average naturalness of 2.6. The SAT-adapted models display an average target speaker similarity of 1.9, which is slightly inferior to the equivalent standard SI models. So no evidence has been observed in this evaluation to support the hypothesis that the SAT estimation technique yields superior models to the standard SI estimation method depicted in Figure 3.

5) Does the constrained two-pass decision tree construction process affect the quality of the resulting SAT-estimated models?: The SAT-estimated models corresponding to standard decision tree construction (system H) yield a slightly superior average naturalness (2.9) to the SAT-estimated models which deploy two-pass decision tree construction (system I, 2.8). As in the case for standard SI models (Section VII-D), constraining decision tree construction using the two-pass technique has slightly compromised the naturalness of the SAT-estimated models.

IX. Conclusion

This paper has introduced a two-pass decision tree construction method. This method enables sharing between full context models used for HMM-based speech synthesis and triphone models used for HMM-based ASR via a simple mapping between these models. This, in turn, enables unsupervised intralingual adaptation of speech synthesis models without a separately estimated set of components. Further, the technique enables the components to be adapted without the use of linguistic analysis. A cross-lingual adaptation technique which uses no source language knowledge is then proposed. This method is based upon the unsupervised intralingual adaptation method. Listener evaluations demonstrate that the proposed unsupervised adaptation methods, both intralingual and cross-lingual, deliver performance approaching that of supervised adaptation.

Several lines of potential future research are directly linked to this work. With regard to the results of Section VII-E, the relationship between the quality of the estimated transcription of the adaptation data and the optimal choice of adaptation algorithm (full or triphone adaptation) merits further investigation.

Future work may also address the reasonably large reductions in naturalness which are observed in the adapted systems of Section VIII-D (in comparison to the unadapted systems).

As mentioned in Section VIII-E, further analysis is required to explain the varying effectiveness of unsupervised cross-lingual adaptation. The relationship between this effectiveness and, for example, adaptation data content or speaker and language characteristics, remains unknown.

Lastly, future work may evaluate the effectiveness of cross-lingual adaptation in the context of an application, for example a personalised speech-to-speech translation system.

X. Acknowledgements

The authors are very grateful to the organizers of the Blizzard Challenge for providing experimental evaluation scripts. This research was partially funded by the European Communitys Seventh Framework Programme (FP7/2007-2013), grant agreement 213845 (EMIME).

References

[1] V. Karaiskos, S. King, R. Clark, and C. Mayo, "The Blizzard Challenge 2008," in *Proc. Blizzard 2008*, 2008.

[2] T. Yoshimura, T. Masuko, K. Tokuda, T. Kobayashi, and T. Kitamura, "Speaker interpolation for HMM-based speech synthesis system," *The Journal of the Acoustical Society of Japan*, vol. 21, no. 4, pp. 119–206, 2000.

[3] J. Yamagishi, K. Onishi, T. Masuko, and T. Kobayashi, "Acoustic modeling of speaking styles and emotional expressions in HMM-based speech synthesis," *IEICE Trans. Information and Systems*, vol. E88-D, no. 3, pp. 503–509, 2005.

[4] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Audio, Speech & Language Processing*, vol. 17(1), pp. 66–83, 2009.

[5] S. King, K. Tokuda, H. Zen, and J. Yamagishi, "Unsupervised adaptation for HMM-based speech synthesis," in *Proceedings Interspeech*, 2008.

[6] B. Wheatley, K. Kondo, W. Anderson, and Y. Muthusamy, "An evaluation of cross-language adaptation for rapid HMM development in a new language," in *Proceedings ICASSP*, vol. 1, 1994, pp. 237–240.

[7] P. Fung, C. Y. Ma, and W. K. Liu, "MAP-based cross-language adaptation augmented by linguistic knowledge: from English to Chinese," in *Proceedings Eurospeech*, 1999, pp. 871–874.

[8] Y. Wu, S. King, and K. Tokuda, "Cross-lingual speaker adaptation for HMM-based speech synthesis," in *Proceedings ISCSLP*, 2008.

[9] Y. Wu, Y. Nankaku, and K. Tokuda, "State mapping based method for cross-lingual speaker adaptation in HMM-based speech synthesis," in *Proceedings Interspeech*, 2009.

[10] J. Yamagishi, H. Zen, Y.-J. Wu, T. Toda, and T. Tokuda, "The HTS-2008 system: yet another evaluation of the speaker-adaptive HMM-based speech synthesis system in the 2008 Blizzard Challenge," in *Proceedings Blizzard*, 2008.

[11] M. Gibson, "Efficient maximum likelihood linear regression," MPhil thesis, Cambridge University, 2004.

[12] J. Dines, L. Saheer, and H. Liang, "Speech recognition with speech synthesis models by marginalising over decision tree leaves," in *Proceedings Interspeech*, 2009.

[13] M. Gibson, "Two-pass decision tree construction for unsupervised adaptation of HMM-based synthesis models," in *Proceedings Interspeech*, 2009.

[14] H. Zen, T. Toda, M. Nakamura, and K. Tokuda, "Details of Nitech HMM-based speech synthesis system for the Blizzard Challenge 2005," *IEICE Trans. Information and Systems*, vol. E90-D, no. 1, pp. 325–333, 2007.

[15] M. Gibson, T. Hirsimaki, R. Karhila, M. Kurimo, and W. Byrne, "Unsupervised cross-lingual speaker adaptation for HMM-based speech synthesis using two-pass decision tree construction," in *Proceedings ICASSP*, 2010.

[16] T. Anastasakos, J. McDonough, R. Schwartz, and J. Makhoul, "A compact model for speaker adaptive training," in *Proceedings ICSLP*, Philadelphia, 1996, pp. 1137–1140.

[17] J. Yamagishi, "Average-voice-based speech synthesis," Ph.D. dissertation, Tokyo Institute of Technology, 2006.

[18] H. Kawahara, I. Masuda-Katsuse, and A. Cheveigne, "Restructuring speech representations using a pitch-adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: possible role of a repetitive structure in sounds," *Speech Communication*, vol. 27, pp. 187–207, 1999.

[19] K. Tokuda, T. Masuko, N. Miyazaki, and T. Kobayashi, "Multi-space probability distribution HMM," *IEICE Trans. Inf. & Syst.*, vol. E85-D(3), pp. 455–464, 2002.

[20] H. Zen, K. Tokuda, T. Masuko, K. Kobayashi, and T. Kitamura, "Hidden semi-Markov model based speech synthesis system," *IEICE Trans. Information and Systems*, vol. E90-D, no. 5, pp. 825–834, 2007.

[21] K. Tokuda, T. Yoshimura, T. Masuko, T. Kobayashi, and T. Kitamura, "Speech parameter generation algorithms for HMM-based speech synthesis," in *Proceedings ICASSP*, 2000.

[22] T. Toda and K. Tokuda, "A speech parameter generation algorithm considering global variance for HMM-based speech synthesis," *IEICE Transactions*, vol. E90-D, no. 5, pp. 816–824, 2007.

[23] J. Yamagishi, B. Usabaev, S. King, O. Watts, J. Dines, J. Tian, R. Hu, Y. Guan, K. Oura, K. Tokuda, R. Karhila, and M. Kurimo, "Thousands of voices for HMM-based speech synthesis," in *Proc. Interspeech*, 2009, pp. 420–423.

[24] V. Digalakis, D. Rtischev, and L. Neumeyer, "Speaker adaptation using constrained reestimation of Gaussian mixtures," *IEEE Transactions Speech and Audio Processing*, pp. 357–366, 1995.

[25] R. Likert, "A technique for the measurement of attitudes," *Archives of Psychology*, vol. 140, pp. 1–55, 1932.

[26] R. Clark, M. Podsiadlo, M. Fraser, C. Mayo, and S. King, "Statistical analysis of the Blizzard Challenge 2007 listening test results," in *Proceedings Blizzard*, 2007.

[27] P. Koehn, "Europarl: A Parallel Corpus for Statistical Machine Translation," in *MT Summit*, 2005.

Matthew Gibson received the B.Sc(Hons) in 1994 from Glasgow University, M.Sc. in 1996 from Oxford University, M.Phil. in 2004 from Cambridge University and PhD in 2008 from Sheffield University. He is currently a research associate at Cambridge University engineering department. His main research interests are machine learning, automatic speech recognition and speech synthesis.

William Byrne (M'82, SM'07) is a Reader in Information Engineering in the Department of Engineering, University of Cambridge. He received his PhD in electrical engineering from the University of Maryland, College Park in 1993 after which he joined The Johns Hopkins University Center for Language and Speech Processing as an Associate Research Scientist and then as Research Associate Professor. He has worked with several speech technology companies, including Entropic Research Laboratory and Voice Signal Technology. His current research interests are in speech recognition, speech synthesis, and statistical machine translation. He is a Fellow of Clare College, Cambridge.

Dr. Byrne was general co-Chair for the 2003 IEEE Automatic Speech Recognition and Understanding Workshop, a member of the Speech Technical Committee from 2004 through 2006, and an Associate Editor of these Transactions from 2006-2008.