A deep learning approach to assessing non-native pronunciation of English using phone distances

Konstantinos Kyriakopoulos, Kate M. Knill, Mark J.F. Gales

ALTA Institute / Engineering Department
Cambridge University
Trumpington St, Cambridge CB2 1PZ, UK
{kk492, kate.knill, mjfg}@eng.cam.ac.uk

Abstract

The way a non-native speaker pronounces the phones of a language is an important predictor of their proficiency. In grading spontaneous speech, the pairwise distances between generative statistical models trained on each phone have been shown to be powerful features. This paper presents a deep learning alternative to model-based phone distances in the form of a tunable Siamese network feature extractor to extract distance metrics directly from the audio frame sequence. Features are extracted at the phone instance level and combined to phone-level representations using an attention mechanism. Pair-wise distances between phone features are then projected through a feed-forward layer to predict score. The extraction stage is initialised on either a binary phone instance-pair classification task, or to mimic the model-based features, then the whole system is fine-tuned end-to-end, optimising the learning of the distance metric to the score prediction task. This method is therefore more adaptable and more sensitive to phone instance level phenomena. Its performance is compared against a DNN trained on Gaussian phone model distance features.

Index Terms: pronunciation assessment, phone distances, CALL, CAPT, Siamese Networks, attention mechanism

1. Introduction

The growing global demand for foreign language learning [1], combined with recent advances in computing power, speech processing and machine learning, has driven an increased interest in Computer Assisted Language Learning (CALL) and auto-marking, in particular the automatic assessment of non-native speaker proficiency [2, 3, 4].

Pronunciation is a key predictor of proficiency, and is expected to become more native, reducing strain to the listener caused by L1 effects, as the learner progresses up the CEFR levels [5]. There is a broad literature on the definition of good pronunciation as it pertains to automatic assessment, with variations in terminology. In this work, pronunciation is defined as the manner in which each word of an utterance is rendered as series of phones, distinguishing it from message construction, which relates to the choice of words, and prosody, which relates to other acoustic properties of speech (i.e. tempo, rhythm and stress). Good pronunciation consists of knowing the correct phone sequence for each word and then rendering those phones in an acceptable manner. A speaker can thus be assessed by the frequency in their speech of *lexical errors* (e.g. pronouncing the silent b in subtle), or by the general way in which they pronounce the phones of the language (e.g. consistently mispronouncing /v/ as /b/). It is this latter accent quality factor, represented by a proficiency score, that this paper concerns.

Section 2 provides an overview of the approaches employed to assess pronunciation in the literature and explains the method of model-based phone distance features, used here as a baseline. Section 3 demonstrates how the phone distance concept can be expanded upon by replacing the model-based approach with a deep, tunable feature extractor based on Siamese networks, while section 4 shows how these can be integrated in an end-to-end system to predict grade. Section 5 presents the data and speech recognition system used in the experiments, while Sections 6 and 7 present the results and conclusions.

2. Phone distance features

Approaches in the literature to pronunciation assessment include comparison to native speaker models [6, 7, 8] and automatic speech recogniser (ASR) confidence measures (usually on alignment tasks) such as Goodness of Pronunciation (GOP) [9, 10, 7, 6]. The problem with both of these approaches is that they generally require prior knowledge of the exact text the speaker is saying (to identify comparable native models in the former case and to give meaning to the ASR confidence scores in the latter). For this reason, most systems in the literature rely on ”read aloud” tasks with known transcriptions. Open responses to questions, however, give a better indication of the learner's proficiency than read speech. When dealing with the resulting spontaneous speech the candidate's audio must first be passed through an ASR, to determine what the speaker said. The recognised text is then used together with the audio for feature extraction to form the input to the automatic grader. This introduces the problem of the ASR's word and phone error rates, which can be particularly high when assessing low proficiency non-native speakers, and to which any assessment system must therefore be robust.

To overcome these issues, more recent approaches to pronunciation assessment have utilised the distances between phones [11, 12, 13, 14]. Rather than characterising each phone by the distribution of acoustic features in its articulations, each phone is defined relative to the pronunciation of each of the others (Figure 1).

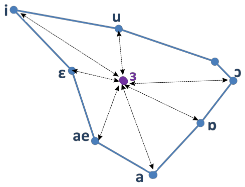

Figure 1: *Illustration of the phone distance concept*

The full set of phone-pair distances describes the speaker's overall accent. These features should thus robustly represent pronunciation in a way that is compact and independent of speaker attributes. Kyriakopoulos et al. [13] proposed a model-based phone distance feature system for pronunciation assessment. Gaussian models are built for each phone and the K-L divergences between them used to determine the speaker's accent quality. This system is taken as the baseline in this paper.

Consider a speaker n whose utterances have been recognised and time-aligned to a series of I_n phone instances $\pi_0, \pi_1, ..., \pi_{I_n - 1}$, each corresponding to one of the 47 English phones $\omega_0, \omega_1, ..., \omega_{46}$ (/aa/, /ax/, /ah/ etc. [1]).

Each phone instance π_i is itself a sequence of T_i frames $\boldsymbol{x}_0^{(i)}, \boldsymbol{x}_1^{(i)}, ..., \boldsymbol{x}_{T_i-1}^{(i)}$, where $\boldsymbol{x}_t^{(i)}$ is the vector of perceptual linear prediction (PLP) features extracted from frame t. Each phone ω_ϕ is represented by the parameters of a multivariate Gaussian model ($\boldsymbol{\mu}_\phi^{(n)}$,$\boldsymbol{\Sigma}_\phi^{(n)}$), which predicts the PLP features produced each time the speaker utters an instance of:

$$p(\boldsymbol{x}_t^{(i)} | \pi_i = \omega_\phi) = \mathcal{N}(\boldsymbol{x}_t^{(i)}; \boldsymbol{\mu}_\phi^{(n)}, \boldsymbol{\Sigma}_\phi^{(n)}) \tag{1}$$

Distances between models are represented using symmetric KullbackLeibler (K-L) divergence:

$$\begin{aligned} D_{\phi,\psi}^{(n)} = \frac{1}{2} \langle \mathcal{KL} \left(\mathcal{N}(\boldsymbol{\mu}_\phi^{(n)}, \boldsymbol{\Sigma}_\phi^{(n)}) || \mathcal{N}(\boldsymbol{\mu}_\psi^{(n)}, \boldsymbol{\Sigma}_\psi^{(n)}) \right) \\ + \mathcal{KL} \left(\mathcal{N}(\boldsymbol{\mu}_\psi^{(n)}, \boldsymbol{\Sigma}_\psi^{(n)}) || \mathcal{N}(\boldsymbol{\mu}_\phi^{(n)}, \boldsymbol{\Sigma}_\phi^{(n)}) \right) \rangle \end{aligned} \tag{2}$$

Each speaker is thus represented by 1081 scalar phone-pair distances $D_{0,1}^{(n)}, D_{0,2}^{(n)}, ..., D_{46,47}^{(n)}$, together forming the vector $\boldsymbol{D}_n$. These features are then used to train a deep neural network (DNN) to predict human-assigned proficiency scores:

$$s_n = f(\boldsymbol{D}_n) \tag{3}$$

3. Siamese phone distance features

This paper proposes an alternative phone distance feature approach to the model-based method described in the previous section. The generative Gaussian model of each phone is replaced with a feature vector projected up from the frame sequences. This feature extractor is tunable and can be fine tuned to the task of automatic assessment, unlike the more general model-based approach. This also addresses a couple of issues with the model-based phone distance feature approach above.

By localising the phone representation to the level of the individual phone instance the amount of data needed to extract the features is reduced. In addition, the features should be more interpretable. The first step is to project the frame sequence $\boldsymbol{x}_0^{(i)}, \boldsymbol{x}_1^{(i)}, ..., \boldsymbol{x}_{T_i-1}^{(i)}$ of each phone instance i to a fixed-length vector representation $\boldsymbol{h}_i$ by passing it through a bi-directional Long Short Term Memory (LSTM) network:

$$\boldsymbol{h}_t^{(f,i)} = f(\boldsymbol{x}_t^{(i)}, \boldsymbol{h}_{t-1}^{(f,i)}, \boldsymbol{\lambda}^{(f,i)}) \tag{4}$$

$$\boldsymbol{h}_t^{(b,i)} = f(\boldsymbol{x}_t^{(i)}, \boldsymbol{h}_{t+1}^{(b,i)}, \boldsymbol{\lambda}^{(b,i)}) \tag{5}$$

The standard mechanism of obtaining a fixed-length vector $\boldsymbol{h}_i$ from the forward and backward sequences of a Recurrent Neural Network (RNN) is to concatenate the two sequences' final time steps (Figure 2, left):

$$\boldsymbol{h_i} = [\boldsymbol{h}_{T-1}^{(f,i)T},\ \boldsymbol{h}_0^{(b,i)T}]^T \tag{6}$$

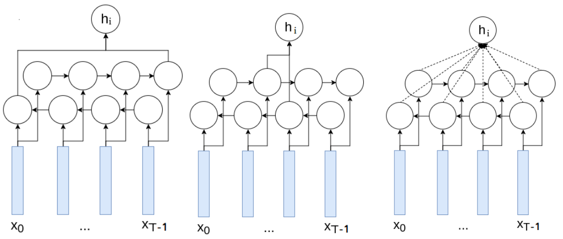

Figure 2: *Standard, centre and attention-based sequence-to-vector bi-RNNs.*

This approach is problematic in this case, however, since the first and last frames of each phone instance are actually the least representative of its content (since they are on the alignment boundary). Two alternative methods are therefore considered:

1. Using the middle, rather than final, frame of each pass (Figure 2, middle)

$$\boldsymbol{h}_i = [\boldsymbol{h}_{T/2}^{(f,i)T},\ \boldsymbol{h}_{T/2}^{(b,i)T}]^T \tag{7}$$

2. Using an attention mechanism to weight the importance of each frame (Figure 2, right)

$$\boldsymbol{h_i} = \sum_{t=0}^{T_i-1} \alpha_t^{(f,i)} \boldsymbol{h}_t^{(f,i)} + \sum_{t=0}^{T_i-1} \alpha_t^{(b,i)} \boldsymbol{h}_t^{(b,i)} \tag{8}$$

where

$$\alpha_t^{(f,i)} = \frac{\exp c_t^{(f,i)}}{\sum_{t=0}^{T_i-1} \exp c_t^{(f,i)} + \exp c_t^{(b,i)}} \tag{9}$$

$$c_t^{(f,i)} = f(\boldsymbol{h}_t^{(f,i)}, \boldsymbol{\lambda}_\alpha) \tag{10}$$

and similarly for $\alpha_t^{(b,i)}$ (with same $\boldsymbol{\lambda}_\alpha$).

Whichever of the three methods is used, the resultant network must be trained across all instances across all speakers to map frame sequences to a space in which distances between phones are most indicative of proficiency.

Having defined a projection from the original audio frame sequences to a phone instance feature space, it is now necessary to use these projections to learn a phone distance metric. To this end, Siamese networks are employed. A Siamese network is composed of two copies of the same neural network, each fed with one of the elements of a pair of input samples. These identical networks project the samples into an embedding space. A measure of distance is then computed between the two samples depending on their relation label (same or different). The error is propagated evenly in the two copies. The architecture is based on the LSTM Siamese architecture for learning difference metrics between pairs of variable length sequences presented by Mueller and Thyagarajan[16].

Consider a pair of phone instances π_i and π_j, taken from the same speaker n, of phones ω_ϕ and ω_ψ respectively. Using the method from the previous section, they are projected to vectors $\boldsymbol{h}_i$ and $\boldsymbol{h}_j$. The distance between the two instances can therefore be represented as:

$$d_{i,j} = ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2 \tag{11}$$

[1] based on ARPABET phone set [15]

Two approaches are considered for training this metric to approximate the distance between the two instances. The first is simply to train $d_{i,j}$ to approximate the model-based K-L divergence distance $D^{(n)}_{\phi,\psi}$ for phones ϕ and ψ in speaker n. The second is to use $d_{i,j}$ to predict whether the two instances are instances of the same phone:

$$c_{ij} = \left\{ \begin{array}{ll} 1, & \omega_\phi = \omega_\psi \\ 0 & \omega_\phi \neq \omega_\psi \end{array} \right. \tag{12}$$

This is done by passing $d_{i,j}$ through a sigmoid to derive the probability p_{ij} that $c_{ij} = 1$

$$p_{ij} = \frac{2}{1 + \exp(-d_{ij})} - 1 \tag{13}$$

Given the large number of possible instance pairs for each speaker n, M pairs $(l^{(n)}_0, r^{(n)}_0), (l^{(n)}_1, r^{(n)}_1), ..., (l^{(n)}_{M-1}, r^{(n)}_{M-1})$ are sampled for use in training (in the experiments in this paper $M = 100$). In the K-L training case sampling is completely random, while in the binary case $\frac{M}{2}$ pairs of instances of the same phone and $\frac{M}{2}$ pairs of instances of different phones are sampled.

If the M pairs have corresponding phone labels $(\phi^{(n)}_0, \psi^{(n)}_0), (\phi^{(n)}_1, \psi^{(n)}_1), ..., (\phi^{(n)}_{M-1}, \psi^{(n)}_{M-1})$, the objective functions over all N speakers in the training becomes, for K-L training:

$$\min \left\{ \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} log|d_{l^{(n)}_m, r^{(n)}_m} - D^{(n)}_{\phi^{(n)}_m, \psi^{(n)}_m} \right\} \tag{14}$$

and for binary training:

$$\min \sum_{n=0}^{N-1} \sum_{m=0}^{M-1} \left\{ \begin{array}{ll} log(p_{l^{(n)}_m, r^{(n)}_m}), & \phi^{(n)}_m = \psi^{(n)}_m \\ log(1 - p_{l^{(n)}_m, r^{(n)}_m}) & \phi^{(n)}_m \neq \psi^{(n)}_m \end{array} \right. \tag{15}$$

The vector $\boldsymbol{h}_i$ for each instance i can now be said to represent a space in which Euclidean distance is representative of the conceptual distance between phones. This representation must now be combined from the phone instance to the overall phone level for each speaker and finally projected to predict speaker score.

4. Predicting grade

Once trained, the systems described in the previous section can extract distance features at the level of phone instance pairs. The next step is to integrate these systems into an end-to-end system for predicting a speaker's proficiency score.

The projected phone instance features $\boldsymbol{h}^{(n)}_i$ for all instances $\pi^{(n)}_i$ of phone $\omega^{(n)}_\phi$ in speaker n, can be combined to derive an overall phone vector $\boldsymbol{h}^{(n)}_\phi$ in one of two ways:

1. By simple averaging:

$$\boldsymbol{h}^{(n)}_\phi = \frac{1}{\sum_{i:\pi_i = \omega_\phi} 1} \sum_{i:\pi_i = \omega_\phi} \boldsymbol{h}^{(n)}_i \tag{16}$$

2. Via another attention mechanism:

$$\boldsymbol{h}^{(n)}_\phi = \sum_{i:\pi_i = \omega_\phi} \alpha^{(n)}_i \boldsymbol{h}^{(n)}_i \tag{17}$$

$$\alpha^{(n)}_i = \frac{\exp c^{(n)}_i}{\sum_{i:\pi_i = \omega_\phi} \exp c^{(n)}_i} \tag{18}$$

$$c^{(n)}_i = g(\boldsymbol{h}^{(n)}_i) \tag{19}$$

where g is a word-level LSTM, since each instance's surrounding context is expected to affect how important it is to score.

The distance feature between each pair of phones ω_ϕ, ω_ψ can now be calculated as:

$$d^{(n)}_{\phi,\psi} = ||\boldsymbol{h}_\phi - \boldsymbol{h}_\psi||_2 \tag{20}$$

resulting in 1081 scalar distances $d^{(n)}_{0,1}, d^{(n)}_{0,2}, ..., d^{(n)}_{46,47}$, together forming vector $\boldsymbol{d}_n$ which is passed through a feed-forward layer to predict the score:

$$s_n = f(\boldsymbol{d}_n) \tag{21}$$

End-to-end training of the whole system (after initialisation using the Siamese network training) with the MSE criterion can now be performed using all N speakers in the training set:

$$\min \left\{ \frac{1}{N} \sum_{n=0}^{N-1} (s_n - f(\boldsymbol{d}_n))^2 \right\} \tag{22}$$

5. Experimental Setup

The preceding sections of this paper have described a system for predicting pronunciation proficiency score based on the audio frames and aligned phone sequence of the utterances produced by a candidate. The data for initialising, training and testing this system are obtained from candidate responses to the spoken component of the Business Language Testing Service (BULATS) for foreign learners of English, provided by Cambridge English Language Assessment. The BULATS speaking test has five sections, all related to business scenarios [17]. Section A consists of short responses to prompted questions. Candidates read 8 sentences aloud in Section B. Sections C-E consist of spontaneous responses of several sentences in length to a series of spoken and visual prompts. Candidates are scored on a scale from 0 to 30, based on their overall proficiency, and it is this score that the system is predicting.

The systems are trained on a gender and proficiency level balanced mixed L1 dataset (TRN) consisting of 994 speakers (first languages Polish, Vietnamese, Arabic, Dutch, French and Thai), scored on their overall proficiency (not just pronunciation) by human graders and evaluated on a held out evaluation set (EVL), consisting of 226 speakers of a similar mix of L1s, gender and proficiency, with scores provided by expert human graders.

As discussed in Section 2, the first step before passing the date through the system is recognising the text being spoken and aligning the audio to a sequence of phones. Both these tasks are performed using an automatic speech recogniser (ASR). Due to the incorrect pronunciations, grammar and rhythm, related to the speaker's proficiency level and first language (L1), the accuracy of standard commercial ”off-the-shelf” ASR systems is too low for non-native learner English. Instead, the ASR system from Kyriakopoulos et al. [13] (also described in Van Dalen et al.[18]), which is trained on non-native learners of English, is used. This ASR has an overall word error rate (WER) of 47.5% and a phone error rate (PER) of 33.9%, evaluated against crowd sourced transcriptions of EVL.

6. Experimental Results

Sections 3 and 4 describe a pronunciation assessment system in two stages. First, an LSTM of either the standard, centre or attention variety is trained, in a Siamese architecture, to extract phone instance features to predict, for a given phone instance pair, either a binary or a K-L divergence distance metric. Next, the trained LSTM is integrated into an end-to-end score prediction system, using either an averaging or attention mechanism to move from the phone instance level to the phone level.

Given the above, it is necessary to evaluate, using the setup described in Section 5, first, how well the system performs at the initialisation tasks, second, which of the proposed architectures (standard vs. centre vs. attention LSTM, binary vs. K-L training, averaging vs. attention combination) is best performing and, third, how the best architecture performs relative to the baseline.

First, the Siamese networks presented in Section 3 are trained on randomly selected pairs from all the speakers in the TRN data set. They are evaluated on similar pairs from the EVL data set for the two tasks of binary (same vs. different) classification and predicting the K-L divergences from the baseline model. Table 1 shows the results for the standard bi-LSTM configuration where the final time steps are concatenated to form the fixed length phone instance representation $\boldsymbol{h}_i$.

Criterion	Binary Accuracy	K-L Performance
Binary	75.0%	0.599
K-L	68.0%	0.789

Table 1: *Performance of standard bi-LSTM configuration Siamese networks*

As expected, the networks perform better on the task they are trained for than on the other task. Both systems perform well, suggesting that the Siamese networks are capable of extracting interpretable distance metrics indicative of both the clustering together of instances of the same phone as well as the distances between distributions of different phones. Further, the fact that the system trained for each task also performs reasonably on the other task, suggests that these two concepts of distance are closely related, as was to be expected.

The experiments are now repeated all three bi-LSTM variants (standard, centre and attention) of the phone instance representation to determine which is best. As expected, and can be seen in Tables 2 and 3, the attention LSTM performs the best overall, with the centre time steps configuration better for matched criteria than the standard configuration. The attention bi-LSTM is slower to train but is used going forward due to this better representation.

Criterion	Std	Cen	Att
Binary	75.0%	77.5%	77.3%
K-L	68.0%	67.4%	69.0%

Table 2: *Binary accuracy of standard (Std), centre (Cen) and attention (Att) bi-LSTM Siamese network configurations*

Criterion	Std	Cen	Att
Binary	0.599	0.587	0.602
K-L	0.789	0.792	0.788

Table 3: *K-L performance of standard (Std), centre (Cen) and attention (Att) bi-LSTM Siamese network configurations*

Having established that the Siamese networks seem to indeed be extracting valid distance features, these features are now employed to predict proficiency scores. The system is connected end-to-end using the averaging and attention methods and further trained to predict grade. The mean squared error (MSE) results of these experiments are presented in Table 4.

Initialisation	MSE	
	Average	Attention
Binary	19.7	17.6
K-L	16.4	14.2

Table 4: *Performance (mean squared error of predicted to human-assigned scores) of baseline and Siamese systems, trained using binary and K-L divergence criteria, with or without the extra attention layer and fine tuning, each trained on TRN and evaluated on EVL*

All systems are able to predict score with a reasonable amount of accuracy and, as expected, adding the attention mechanism improves performance. The systems which were initialised using the K-L divergences from the baseline method outperform those initialised using the binary classifier, which is to be expected given the superior granularity of K-L divergences compared to the binary variant.

Finally, the performance of the best architecture (attention LSTM, K-L criterion, attention combination), is compared to that of the baseline (Table 5). The system outperforms the baseline in terms of MSE but is comparable for PCC. This can be explained given that the new system is optimised end-to-end for minimum MSE, whereas for the baseline only the grader is optimised for minimum MSE.

Initialisation	MSE	PCC
Baseline	14.8	0.785
End-to-end system	14.2	0.780

Table 5: *Performance (mean squared error and Pearson correlation coefficient of predicted to human-assigned scores) of baseline and Siamese systems, trained using binary and K-L divergence criteria, each trained on TRN and evaluated on EVL*

7. Conclusions

Phone distance features have previously been shown to be a good indicator of accent pronunciation quality allowing use in assessing the proficiency of a non-native learner's speech, in particular, for assessment of spontaneous spoken responses. This paper has proposed an alternative to the model-based approaches to phone distances based on Siamese networks.

It was first shown how Siamese networks can be used to extract distance metrics between pairs of phone instances. These can be used to predict whether the two are instances of the same or different phones, as well as the relative entropies between the distributions of their two phones if they are instances of different phones. The distance measure is tunable, performing differently on different tasks depending on how it is trained. The latter paradigm, which involves calculating model-based relative entropies before training, was shown to produce a superior distance measure. Of the three different architectures considered, that involving an attention mechanism outperformed the standard and centre-based sequence-to-vector architectures.

Finally, the networks were also used to develop a proficiency grader, alternatively using an averaging and attention mechanism to move from the phone instance level to the phone level, with the latter proving superior. Trained in an end-to-end fashion this grader was able to predict human-assigned proficiency scores with performance surpassing model-based phone distance features.

8. References

[1] D. Graddol, *English Next*. British Council, 2006.

[2] S. M. Witt, "Automatic error detection in pronunciation training: Where we are and where we need to go," in *Proc. IS ADEPT*, vol. 6, 2012.

[3] S. Detey, L. Fontan, and T. Pellegrini, "Traitement de la prononciation en langue étrangère: approches didactiques, méthodes automatiques et enjeux pour lapprentissage," *Revue Traitement Automatique des Langues*, 2016.

[4] R. van Dalen, K. Knill, and M. Gales, "Automatically Grading Learners' English Using a Gaussian Process," in *SLaTE*, 2015.

[5] Council of Europe, *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.

[6] N. Moustroufas and V. Digalakis, "Automatic pronunciation evaluation of foreign speakers using unknown text," *CSL*, vol. 21, pp. 219–230, 2007.

[7] A. Metallinou and J. Cheng, "Using Deep Neural Networks to Improve Proficiency Assessment for Children English Language Learners," in *INTERSPEECH*, 2014, pp. 1468–1472.

[8] M. Nicolao, A. V. Beeston, and T. Hain, "Automatic assessment of english learner pronunciation using discriminative classifiers," in *ICASSP*. IEEE, 2015, pp. 5351–5355.

[9] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *SPEECHCOM*, vol. 30, pp. 95–108, 2000.

[10] H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen, "Combination of machine scores for automatic grading of pronunciation quality," *SPEECHCOM*, 2000.

[11] N. Minematsu, S. Asakawa, and K. Hirose, "Structural representation of the pronunciation and its use for CALL," in *SLT*, 2006, pp. 126–129.

[12] S. Kasahara, N. Minematsu, H. Shen, D. Saito, and K. Hirose, "Structure-based prediction of english pronunciation distances and its analytical investigation," in *Information Science and Technology (ICIST), 2014 4th IEEE International Conference on*. IEEE, 2014, pp. 332–336.

[13] K. Kyriakopoulos, M. Gales, and K. Knill, "Automatic Characterisation of the Pronunciation of Non-native English Speakers using Phone Distance Features," in *SLATE*, 2017.

[14] K. Knill, M. Gales, K. Kyriakopoulos, A. Ragni, and Y. Wang, "Use of Graphemic Lexicons for Spoken Language Assessment," in *INTERSPEECH*, 2017, pp. 2774–2778.

[15] D. Jurafsky and J. H. Martin, *Speech and Language Processing*. New Jersey: Prentice-Hall, 2000.

[16] J. Mueller and A. Thyagarajan, "Siamese Recurrent Architectures for Learning Sentence Similarity," in *AAAI*, 2016, pp. 2786–2792.

[17] L. Chambers and K. Ingham, "The BULATS online speaking test," *Research Notes*, vol. 43, pp. 21–25, 2011. [Online]. Available: http://www.cambridgeenglish.org/images/23161-research-notes-43.pdf

[18] R. C. van Dalen, K. M. Knill, P. Tsiakoulis, and M. J. F. Gales, "Improving multiple-crowd-sourced transcriptions using a speech recogniser," in *ICASSP*, Apr 2015.