HMM-based Non-native Accent Assessment using Posterior Features

Ramya Rasipuram, Milos Cernak and Mathew Magimai-Doss

Idiap Research Institute, Martigny, Switzerland

ramya.rasipuram@gmail.com, milos.cernak@idiap.ch, mathew@idiap.ch

Abstract

Automatic non-native accent assessment has potential benefits in language learning and speech technologies. The three fundamental challenges in automatic accent assessment are to characterize, model and assess individual variation in speech of the non-native speaker. In our recent work, accentedness score was automatically obtained by comparing two phone probability sequences obtained through instances of non-native and native speech. Although automatic accentedness ratings of the approach correlated well with human accent ratings, the approach is critically constrained because of the requirement of native speech instance. In this paper, we build on the previous work and obtain the native latent symbol probability sequence through the word hypothesis modeled as a hidden Markov model (HMM). The latent symbols are either context-independent phonemes or clustered context-dependent phonemes. The advantage of the proposed approach is that it requires just reference text transcription instead of native speech recordings. Using the HMMs trained on an auxiliary native speech corpus, the proposed approach achieves a correlation of 0.68 with human accent ratings on the ISLE corpus. This is further interesting considering that the approach does not use any non-native data and human accent ratings at any stage of the system development.

Index Terms: Automatic accent assessment, non-native speech, posterior features, KL-divergence, lexical model

1. Introduction

Automatic accent assessment is an emerging topic of interest in language learning and speech technologies. Non-native accent or foreign accent is characterized by transfer of pronunciation rules, phonetic and prosodic structure from the native language of a speaker to a second language. Accent is typically assessed through perceptual listening tests, where the listeners either assess a particular aspect of accent (for example, phonetic structure or intonation) or general accentedness of a speaker [1, 2]. Accent of a speaker depends on various factors such as age of onset and years of second language learning, language learning aptitude etc. Furthermore, there is also an influence of the listener on perception of non-native accent [2]. Therefore, in the literature there has been a growing interest in fast and reliable automatic accent assessment methods.

Automatic accent assessment could be performed at phone or utterance levels. At the phone level, it is typically formulated as a 2-class classification task to determine if the pronunciation of a phone was correct or not. A variety of confidence measures are extracted at the output of an hidden Markov model (HMM) based speech recognizer such as log-likelihood [3], log-likelihood ratio [4], goodness of pronunciation [5], log-posterior probability scores [3, 6]. Accent assessment approaches based on speech structure [7], phonological features [8, 9], native listener perceptual information [10, 11] etc., have been proposed. Mispronunciation is detected using classifiers such as decision trees [12], logistic regression [13] that combine one or more of the above confidence measures. In [14, 15], a combination of dynamic programming and classifier approaches was proposed for word-level mispronunciation detection. The two main drawbacks of classifier-based approaches are separate classifiers for each phone are needed, and human accent ratings are required.

For utterance level accent evaluation using phonetic structure, phone-level log-likelihood scores were averaged over the utterance [16]. In [17], an intonation-based accent score was obtained through HMMs trained for categorical intonation units. In [1], a large number of rhythm features and prosodic features are used to train a discriminative classifier. In this paper, our interest is in utterance-level accent assessment.

In our recent work [18], we proposed a novel formulation for automatic accent assessment as quantifying the acoustic-phonetic mismatch between latent symbol posterior probability sequences obtained through instances of native and non-native speech. Latent symbols can be context-independent phones or clustered context-dependent phone states. The knowledge of native speech, i.e., the lexical and phonetic structure, was imposed through an instance of native speech. The resulting scores correlated highly with the human accent ratings on English utterances from German, Finnish and Mandarin native speakers.

In this paper, we build upon our previous work along the two following directions (Section 2). Firstly, the lexical and phonetic structure of the native speech are imposed through an HMM-based lexical model trained on native speech data [19]. Specifically, the native reference posterior probability sequence is obtained by modeling the word hypothesis through the Kullback-Leibler divergence based HMM (KL-HMM). Thus the approach is text-independent and it alleviates the need for native reference speech. Secondly, we show that the model-based framework can be exploited to compute confidence measures at various levels. In this paper, word and phone-level confidence measures are computed as the average KL-divergence between the non-native latent symbol and the HMM-based native reference probability sequences.

We evaluate the potential of the approach on the ISLE corpus which contains English speech from native German and Italian speakers (Section 3) [20, 17]. Using HMM models trained on an auxiliary speech corpus and without using any human accent ratings during training, utterance level accent scores computed using the proposed approach correlate well ($R = 0.68$) with the human accent ratings (Section 4).

This research was funded by the Commission for Technology and Innovation (CTI) on "Automatic scoring and adaptive pedagogy for oral language learning (ScoreL2)". The authors would like to thank Prof. Shrikanth Narayanan and Dr. Joseph Tepperman for kindly sharing with us the human accent ratings of the ISLE corpus.

2. Non-native Accent Assessment Approach

In this section, we first briefly eloborate our previous accent assessment approach [18] before presenting the HMM-based accent assessment approach.

2.1. Previous Work

In our recent work, we proposed a novel formulation for automatic accent assessment based on comparison of latent symbol posterior probability sequences obtained through instances of native and non-native speech [18]. The approach is split into four subproblems:

1. **Latent symbols:** The latent symbol set defines the granularity at which the differences between native and non-native speech are captured. In our previous work we showed that the latent symbols can be context-independent phones or clustered context-dependent phone states.
2. **Acoustic model:** The acoustic model, models the relationship between the acoustic feature observations and the latent symbols on native speech data from the target language. As in our previous work, we model this relationship through artificial neural networks (ANNs). Given a non-native speech utterance $X_{nn} = [\mathbf{x}_1^{(nn)}, \cdots, \mathbf{x}_n^{(nn)}, \cdots, \mathbf{x}_N^{(nn)}]$, the acoustic model estimates the latent symbol posterior probability sequence $Z = [\mathbf{z}_1, \cdots, \mathbf{z}_n, \cdots, \mathbf{z}_N]$,

$$\begin{aligned}\mathbf{z}_n &= [z_n^1, \ldots, z_n^k, \ldots, z_n^K]^{\mathrm{T}}, \\ &= [P(c_1|\mathbf{x}_n), \ldots, P(c_k|\mathbf{x}_n), \ldots, P(c_K|\mathbf{x}_n)]^{\mathrm{T}}, \quad (1)\end{aligned}$$

   Here N denotes the number of frames, $c_1, \ldots, c_K$ denote the latent symbols and K denotes the number of latent symbols.
3. **Lexical model:** The lexical model, models the relationship between lexical units (context-dependent subword units) and the latent symbols. In the case of accent assessment, the lexical model imposes the lexical and phonetic structure of native speech utterance. Depending on the way word hypothesis is represented, the lexical model can be instance-based [21] or model-based [19]. In our previous work [18], we focussed on the instance-based lexical model. As shown in Fig 1(a), given the native utterance X_n, latent symbol posterior probability sequence $Y = [\mathbf{y}_1, \cdots, \mathbf{y}_m, \cdots, \mathbf{y}_M]^{\mathrm{T}}$ is estimated using an ANN.
4. **Match between native and non-native sequences:** This matching is typically performed using dynamic programming with local constraints and a local score that matches the acoustic and lexical models evidence at each time frame.

2.2. HMM-based Lexical Modeling for Accent Assessment

In this paper, we build on the approach and obtain the native posterior probability sequence by modeling the word hypothesis through an HMM. As shown in Fig 1(b), the text spoken by the non-native speaker is converted to a sequence of lexical units using a pronunciation lexicon. The sequence of lexical units is represented by a sequence of HMM-states where each HMM-state captures the relationship between lexical unit and latent variables. Each HMM-state is either parameterized by a Kronecker delta distribution (deterministic lexical modeling) or categorical state distribution (probabilistic lexical modeling).

In the case of deterministic lexical modeling, the lexical model, models a deterministic relationship between lexical units and latent symbols. Typically, decision-trees are used to deterministically map each lexical unit to a latent symbol. The decision trees are trained using the pronunciation lexicon, linguistic knowledge (a phonetic question set) and acoustic data of the native speech from the target language. Because of the deterministic relationship between lexical units and latent symbols, the lexical model or the HMM-state distribution is a K-dimensional Kronecker delta distribution. That is $\mathbf{y}_m = [y_m^1, \ldots, y_m^k, \ldots, y_m^K]^{\mathrm{T}}$ and if the lexical unit l_m is mapped to the latent symbol c_j ($l_m \mapsto c_j$) then,

$$y_m^k = \begin{cases} 1, & \text{if k = j;} \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

In the case of the probabilistic lexical modeling, the lexical model captures a probabilistic relationship between lexical units and latent variables. More specifically, the lexical model or HMM-state distribution is a K-dimensional categorical distribution $\mathbf{y}_m = [y_m^1, \ldots, y_m^k, \ldots, y_m^K]^{\mathrm{T}}$ where $y_m^k = P(c_k|l_m)$, $0 < P(c_k|l_m) < 1$ and $\sum_{k=1}^{K} P(c_k|l_m) = 1$. The lexical model parameters are trained on the native speech from the target language using the KL-HMM approach [18].

Match between sequences of native and non-native speech: The non-native latent symbol posterior probability sequence Z is matched with the deterministic or probabilistic lexical model represented by sequence of HMM states through dynamic programming. Specifically, in the case of HMM-based lexical modeling, the Viterbi alignment is used to align the sequences Z and Y using a local score and local HMM constraints.

In the case of deterministic lexical model, the local score that matches the acoustic model evidence $\mathbf{z}_n$ at time frame n with the lexical model evidence $\mathbf{y}_m$ at HMM state m is,

$$S(\mathbf{y}_m, \mathbf{z}_n) = \sum_{k=1}^{K} y_m^k \log\left(\frac{y_m^k}{z_n^k}\right). \quad (3)$$

Since each lexical unit l_m is deterministically mapped to a latent symbol c_j (i.e., $l_m \mapsto c_j$),

$$S(\mathbf{y}_m, \mathbf{z}_n) = -\log P(c_j|q_t = l_m). \quad (4)$$

where q_t is the HMM state at time t. In the case of probabilistic lexical model, the local score matches the posterior distribution $\mathbf{z}_n$ with HMM-state distribution $\mathbf{y}_m$ through the reverse KL-divergence,

$$S(\mathbf{y}_m, \mathbf{z}_n) = \sum_{k=1}^{K} z_n^k \log\left(\frac{z_n^k}{y_m^k}\right). \quad (5)$$

HMM-based lexical modeling provides a framework to compute confidence measures at various levels which can be employed in accent assessment. A confidence measure $C(s_{rm})$ for each phone s_{rm} is computed as the average of the local score between the sequence of posteriors of the non-native speech and the HMM-state distributions i.e.,

$$C(s_{rm}) = \frac{1}{e_{rm} - b_{rm} + 1} \sum_{n=b_{rm}}^{e_{rm}} S(\mathbf{y}_{s_{rm}}, \mathbf{z}_n). \quad (6)$$

Similarly, a confidence measure $C(w_m)$ for each word w_m is computed based on the average of the local score between the sequence of posteriors of the non-native speech and the HMM state distributions,

$$C(w_m) = \frac{1}{R_m} \sum_{r=1}^{R_m} \frac{1}{e_{rm} - b_{rm} + 1} \sum_{n=b_{rm}}^{e_{rm}} S(\mathbf{y}_{s_{rm}}, \mathbf{z}_n). \quad (7)$$

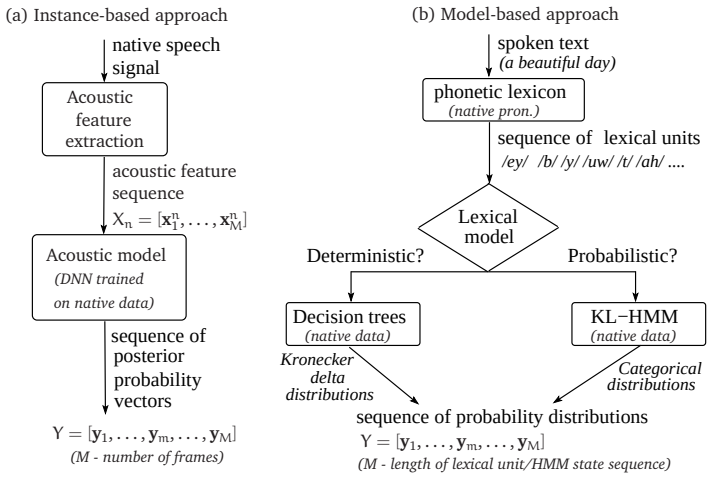

Figure 1: Instance-based and Model-based lexical modeling approaches.

where s_{rm} is the r^{th} subword state in word w_m, b_{rm} and e_{rm} are the begin and end indices of the frames aligned with subword state s_{rm}, and R_m is number of lexical units in word w_m.

The utterance-level accent score is the average of phone-level or word-level confidence measures across the utterance. The phone-level score in the case of deterministic lexical modeling $C(s_{rm})$ is equivalent to the log phone posterior (LPP) of phone as defined in [6]. The deterministic lexical modeling or LPP based accent assessment serves as a baseline for the probabilistic lexical modeling based accent assessment.

3. Experimental Setup

The experimental evaluations presented in this paper are conducted on the data from the ISLE corpus [20]. We used the train and test set division for the ISLE corpus as defined in [17].

Speakers: The study consists of English speech from native German and Italian speakers. The corpus has 8 training speakers and 8 test speakers. The speakers in the train and test sets are different. There are about 150 train and 40 test utterances for each speaker. We did not use the train utterances of the ISLE corpus in our experiments. According to the manual phone-level error labelling [20], native Italian speakers produced more phone errors per word (average of 0.54 errors per word) than the native German (average of 0.16 errors per word). The database did not include reference native speaker utterances.

Human accentedness ratings: We used the human accent ratings collected by the Signal Analysis and Interpretation Laboratory (SAIL) [17]. The sentences were scored taking into account intonation and all other cues on a scale from 1 or "no foreign accent" to 5 "strong foreign accent". Two stage approach was employed to obtain human accent ratings. In the first stage, part of the corpus (138 sentences) was labelled by five native speakers of English. Average inter-labeler correlation of 0.657 was achieved. In the second stage, one native listener who had an average correlation of 0.732 with all the other five listeners, scored all the utterances of the corpus [17].

MLPs: In this paper, we used the same multilayer perceptrons (MLPs) used in our previous study on accent assessment as acoustic models [18]. The MLPs were trained on the Wall Street Journal (WSJ) corpus [22]. The WSJ corpus consists of two parts - WSJ0 with 14 hours of speech (7,193 utterances from 84 speakers) and WSJ1 with 66 hours of speech (29322 utterances from 200 speakers). We used both WSJ0 and WSJ1 (the si-284 setup). We trained the following five-layer MLPs:

- *MLP-CI-40*: An MLP trained to classify 40 context-independent phones.
- *MLP-CD-N*: MLPs trained to classify N context-dependent phone states. The latent symbols or context-dependent phone states were obtained by decision tree-based state clustering of context-dependent phones in HMM/GMM framework. The different number of latent symbols N ($N \in \{183, 419, 1013, 1915, 2832\}$) were obtained by varying the state occupancy count and the log-likelihood threshold during decision-tree based state clustering.

Lexical Model: In the case of the baseline system using deterministic lexical modeling, the accent scores are the same as log phone posterior based accent scores proposed in [6]. The decision trees trained during HMM/GMM training are used to map each context-dependent lexical unit to a latent symbol. The resulting mapping is used to generate Kronecker delta distributions of lexical units. Each lexical unit or context-dependent subword unit was modeled using three HMM-states.

In the case of probabilistic lexical modeling, KL-HMM systems are trained only on the WSJ0 corpus (the si-84 setup) that contains approximately 14 hours of speech. Given the MLPs, first the acoustic unit posterior feature vectors $\mathbf{z}_t$ are estimated for the WSJ0 corpus. The lexical model parameters are then learned using the KL-HMM approach with $\mathbf{z}_t$ as feature observations [18]. We trained crossword context-dependent KL-HMM systems and the lexical units impose three-state minimum duration constraint.

Automatic accentedness evaluation: Utterance-level accent scores are computed using either the phone-level (Eqn. (6)) or word-level confidence measures (Eqn. (7)). Furthermore, the utterance-level accent score is directly correlated (using Pearson correlation coefficient) with the human accent ratings.

4. Results and Analysis

Table 1 presents the utterance level correlation between automatic accent scores computed using phone and word-level confidence measures and human accent ratings for the ISLE test set with increasing phonetic granularity. The results indicate that:

- As the granularity of the latent symbols increases, the correlation with respect to the human ratings generally increases for both deterministic and probabilistic lexical models. This trend was also observed in our previous study on the EMIME corpus using instance-based lexical modeling [18].

Table 1: Correlation between the human accent ratings and the utterance automatic accent scores computed using phone and word-level confidence measures with probabilistic and deterministic lexical models.

# of latent	Probabilistic		Deterministic	
symbols	phone-level	word-level	phone-level	word-level
40	0.58	0.48	0.53	0.40
183	0.63	0.53	0.55	0.40
419	0.66	0.57	0.61	0.50
1013	**0.68**	**0.60**	0.64	0.54
1915	0.67	0.59	0.67	0.55
2832	0.67	0.59	0.67	0.58

- Probabilistic lexical model based systems achieve better correlation than the baseline deterministic lexical model based systems. Furthermore, probabilistic lexical model based system achieved optimal correlation using 1013 latent symbols while the deterministic lexical model based system achieved optimal correlation with 2832 latent symbols. Interestingly, such a trend has also been observed in ASR studies [23].
- The systems using phone-level confidence measures perform better than the systems using word-level confidence measures. This result indicates that phone-level confidence measures are more indicative of the accentedness as perceived by humans than the word-level confidence measures.
- In [17], on the same experimental setup, a correlation of 0.38 with respect to human accent ratings was obtained using prosodic models. In the literature, it has been observed that for advanced language speakers with fluent but accented speech, prosodic-level differences contribute to perceived accent more so than the individual phone mispronunciations, whereas for beginner and intermediate language learners phone level mispronunciations contribute more to the perceived accent [24]. In comparison to prosodic models [17], the proposed approach results in higher correlation with the human ratings. Since the ISLE corpus consists of intermediate learners [20], we speculate that phonetic level assessment performs better than prosodic level assessment.

The results are encouraging given that the approach achieves a correlation of 0.68 without using any non-native data or the human accent ratings during training. To understand the differences among different language groups, we analysed the correlation of native German and Italian utterances separately. Figure 2 plots the correlation achieved with the proposed approach for native German and Italian speakers for both deterministic and probabilistic lexical models using phone-level confidence measures. The plot shows that:

- For native Italians, probabilistic lexical model based systems achieved higher correlation than the baseline deterministic lexical model based systems; whereas for native Germans, deterministic lexical model based systems achieved higher correlation than the probabilistic lexical model based systems. Probabilistic lexical modeling is an approach for pronunciation variability modeling which handles the shortcomings of the deterministic lexical unit to latent symbol modeling of standard HMM-based ASR systems [23, 19]. The results in the paper indicate that for native German speakers whose English is close to the native English speech such pronunciation variability modeling may not be necessary.

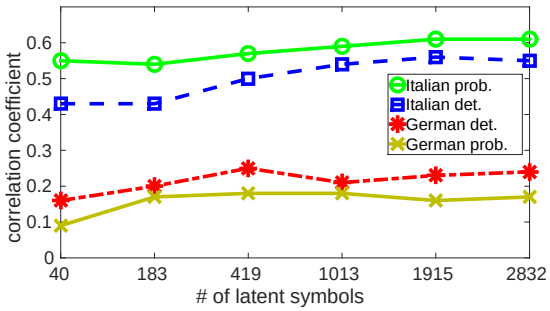

Figure 2: Correlation between human accent ratings and automatic ratings for native German and Italian speakers using the deterministic and probabilistic lexical modeling approaches.

- The correlation between automatic accent ratings and human ratings for native Italian speakers is higher than for the native German speakers. We speculate the following reasons for this: Firstly, it has been observed that it is difficult to rate the accentedness of non-native second language speakers whose speech is closer to the native speech [25, 26]. Secondly, the proposed approach focusses on phone-level (or word-level) mismatch between native and non-native speech. As mentioned in Section 3, according to the manual mispronunciation labels, native German speakers have relatively less phone errors per word compared to native Italian speakers. This leaves less scope for the proposed approach to measure native German speakers accentedness.

5. Conclusions and Future Work

In this paper, we extended our previous work on accent assessment by replacing the native reference posterior probability sequence obtained through an instance of native speech signal with a native posterior probability sequence obtained through an HMM-based lexical model. The HMM-based lexical model requires only the text transcription of the non-native utterance to be assessed and thus removed the constraint that the native reference speech is required. Furthermore, it offered flexibility to compute confidence measures at various levels (word and phone levels) which were used to compute utterance level accent scores. Our studies on the ISLE corpus show that the utterance level accent scores directly correlate well with the human accent ratings. The accent scores based on phone-level confidence measures correlated better with the human accent scores than the scores based on word-level confidence measures. The results are interesting given that the HMM-model was trained on an auxiliary out-of-domain native speech corpus and the approach did not use any non-native speech data or human accent ratings during system development.

Our analysis has shown how native language background of the non-native speakers influences the correlation between automatic accent ratings and human accent ratings. Specifically, we found that for native German speakers (with fewer phone errors per word) the correlation with the human accent ratings is poor compared to native Italian speakers. As indicated in the literature, for advanced non-native speakers, prosodic characteristics may play an important role in accent perception. Therefore, in future we will focus on integrating prosodic characteristics in our formulation (for example, using prosodic representations as given in [17]). Furthermore, we will extend the approach to mispronunciation detection at the phone or word levels by thresholding the confidence measures as done in [27].

6. References

[1] F. Hönig, A. Batliner, and E. Nöth, "Automatic Assessment of Non-Native Prosody - Annotation, Modelling and Evaluation," in *Proceedings of ISADEPT*, 2012.

[2] M. Wester and C. Mayo, "Accent rating by native and non-native listeners," in *Proc. of ICASSP*, 2014, pp. 7699–7703.

[3] Y. Kim, H. Franco, and L. Neumeyer, "Automatic Pronunciation Scoring of Specific Phone Segments for Language Instruction," in *Proc. of EUROSPEECH*, 1997, pp. 64 564–64 568.

[4] H. Franco, L. Neumeyer, M. Ramos, and H. Bratt, "Automatic Detection Of Phone-Level Mispronunciation For Language Learning," in *Proc. of EUROSPEECH*, 1999, pp. 851–854.

[5] S. Witt and S. Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech Communication*, vol. 30, no. 2–3, pp. 95–108, 2000.

[6] W. Hu, Y. Qian, and F. K. Soong, "A new DNN-based high quality pronunciation evaluation for computer-aided language learning (CALL)," in *Proc. of Interspeech*, 2013, pp. 1886–1890.

[7] S. Masayuki, D. Luo, M. Nobuaki, and H. Keikichi, "Improved Structure-based Automatic Estimation of Pronunciation Proficiency," in *SLaTE*, 2009.

[8] A. Sangwan and J. H. Hansen, "Automatic analysis of Mandarin accented English using phonological features," *Speech Communication*, vol. 54, no. 1, pp. 40–54, 2012.

[9] H. Wang, X. Qian, and H. Meng, "Phonological modeling of mispronunciation gradations in L2 English speech of L1 Chinese learners," in *Proc. of ICASSP*, 2014, pp. 7714–7718.

[10] F. William, A. Sangwan, and J. H. Hansen, "Automatic Accent Assessment Using Phonetic Mismatch and Human Perception," *IEEE Trans. on Audio, Speech, and Language Processing*, vol. 21, no. 9, pp. 1818–1829, 2013.

[11] J. Jia, W.-K. Leung, Y.-H. Wu, X.-L. Zhang, H. Wang, L.-H. Cai, and H. M. Meng, "Grading the Severity of Mispronunciations in CAPT Based on Statistical Analysis and Computational Speech Perception," *Journal of Computer Science and Technology*, vol. 29, no. 5, 2014.

[12] A. Ito, Y.-L. Lim, M. Suzuki, and S. Makino, "Pronunciation error detection for computer-assisted language learning system based on error rule clustering using a decision tree," *Acoustical Science and Technology*, vol. 28, no. 2, pp. 131–133, 2007.

[13] W. Hu, Y. Qian, F. K. Soong, and Y. Wang, "Improved mispronunciation detection with deep neural network trained acoustic models and transfer learning based logistic regression classifiers," *Speech Communication*, vol. 67, pp. 154–166, 2015.

[14] A. Lee and J. Glass, "A comparison-based approach to mispronunciation detection," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*, Dec 2012, pp. 382–387.

[15] A. Lee, Y. Zhang, and J. Glass, "Mispronunciation detection via dynamic time warping on deep belief network-based posteriorgrams," in *Proc. of ICASSP*, May 2013, pp. 8227–8231.

[16] H. Li, S. Huang, S. Wang, and B. Xu, "Context-dependent Duration Modelling with Backoff Strategy and Look-up Tables for Pronunciation Assessment and Mispronunciation Detection," in *Proc. of Interspeech*, 2011.

[17] T. Joseph and N. S. S., "Better nonnative intonation scores through prosodic theory," in *Proc. of Interspeech*, 2008, pp. 1813–1816.

[18] R. Rasipuram, M. Cernak, A. Nanchen, and M. Magimai.-Doss, "Automatic accentedness evaluation of non-native speech using phonetic and sub-phonetic posterior probabilities," in *Proc. of Interspeech*, 2015.

[19] R. Rasipuram and M. Magimai.-Doss, "Acoustic and Lexical Resource Constrained ASR using Language-Independent Acoustic Model and Language-Dependent Probabilistic Lexical Model," *Speech Communication*, vol. 68, pp. 23–40, 2015.

[20] W. Menzel, E. Atwell, P. Bonaventura, D. Herron, P. Howarth, R. Morton, and C. Souter, "The ISLE corpus of non-native spoken English," in *Proc. of LREC*, 2000, pp. 957–963.

[21] S. Soldo, M. Magimai.-Doss, J. P. Pinto, and H. Bourlard, "Posterior Features for Template-based ASR," in *Proc. of ICASSP*, 2011.

[22] P. C. Woodland, J. J. Odell, V. Valtchev, and S. J. Young, "Large Vocabulary Continuous Speech Recognition using HTK," in *Proc. of ICASSP*, vol. 2, 1994, pp. 125–128.

[23] M. Razavi, R. Rasipuram, and M. Magimai.-Doss, "On Modeling Context-Dependent Clustered States: Comparing HMM/GMM, Hybrid HMM/ANN and KL-HMM Approaches," in *Proc. of ICASSP*, 2014.

[24] S. M. Witt, "Automatic Error Detection in Pronunciation Training: Where we are and where we need to go," in *International Symposium on automatic detection on errors in pronunciation training*, 2012.

[25] P. Müller, F. D. Wet, C. V. D. Walt, and T. Niesler, "Automatically assessing the oral proficiency of proficient L2 speakers," in *Proceedings of the ISCA Workshop on Speech and Language Technology in Education (SLaTE)*, 2009.

[26] K. Yan and S. Gong, "Pronunciation Proficiency Evaluation based on Discriminatively Refined Acoustic Models," *International Journal of Information Technology and Computer Science (IJITCS)*, vol. 3, no. 2, pp. 17–23, 2011.

[27] R. Ullmann, R. Rasipuram, M. Magimai.-Doss, and H. Bourlard, "Objective Intelligibility Assessment of Text-to-Speech Systems Through Utterance Verification," in *Proc. of Interspeech*, 2015.