Automatic detection of accent and lexical pronunciation errors in spontaneous non-native English speech

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

Detecting individual pronunciation errors and diagnosing pronunciation error tendencies in a language learner based on their speech are important components of computer-aided language learning (CALL). The tasks of error detection and error tendency diagnosis become particularly challenging when the speech in question is spontaneous and particularly given the challenges posed by the inconsistency of human annotation of pronunciation errors. This paper presents an approach to these tasks by distinguishing between lexical errors, wherein the speaker does not know how a particular word is pronounced, and accent errors, wherein the candidate's speech exhibits consistent patterns of phone substitution, deletion and insertion. Three annotated corpora of non-native English speech by speakers of multiple L1s are analysed, the consistency of human annotation investigated and a method presented for detecting individual accent and lexical errors and diagnosing accent error tendencies at the speaker level.

Index Terms: pronunciation, CAPT, CALL, speech recognition

1. Introduction

Computer Assisted Pronunciation Training (CAPT) is an important component of Computer Aided Language Learning (CALL). Two key tasks in providing useful feedback to a learner to improve their pronunciation are error detection, identifying words that are pronounced incorrectly, and error tendency diagnosis, detecting speakers' overall tendencies to make particular types of errors. Spontaneous speech provides additional constraints to CALL tasks, as it tends to be less fluent, containing non-words, grammatical errors and hesitations. The text being spoken is not known in advance and varies between speakers so has to be recognised by an automatic speech recogniser (ASR), inevitably introducing noise in the transcriptions.

The most direct way to evaluate pronunciation is free phone recognition. Systems are trained to directly recognise the apparent phones pronounced by the speaker so they can be compared to canonical pronunciations [1, 2]. Problems with this approach include scarce annotated training data and variability in how different speakers render each phone. These have been tackled by recognising articulatory features instead of phones [3, 4] and employing multi-task learning to incorporate speech data in the speakers' native languages (L1s) [5, 6]. Recently, end-to-end approaches have been developed [7, 8] to directly detect errors from acoustic features.

A common problem, however, underlying all supervised error detection methods is that inter-annotator agreement in labelling of pronunciation errors is difficult to achieve [9]. These agreements have been shown to be better, on the other hand, when diagnosing error tendency at the speaker level [10]. This suggests that, while a speaker's tendency to pronounce a certain phone incorrectly can be ascertained objectively, which of the instances of the speaker's realisations of that phone will be heard as incorrect will vary between listeners.

Approaches to obtain a more reliable ground truth than human annotation include using motion sensors on speakers' mouths [11] and even MRI scanning [12] to directly measure articulation, but such techniques have limited practical applicability due to scarcity of training data and difficulty assigning physical metrics to pronunciation perceived by a listener. Another common approach is dynamic time warping (DTW) [13] to align and compare the non-native speaker's utterance to an utterance of identical text by a native. As this requires prior knowledge of the text being read, however, it cannot be used with spontaneous speech. A related but text-independent method was introduced in [14], using self-DTW to compare tri-phones to other tri-phones within the speech of the same speaker and thus detect errors in a unsupervised manner.

Pronunciation scoring methods detect errors using confidence features derived from the automatic speech recogniser (ASR), including log likelihood [15], likelihood ratio [16, 17], Goodness of Pronunciation (GOP) [18, 6] and phone posterior probability [10, 19]. These approaches are mostly employed on read speech as they rely on the ASR having detected the correct word sequence, though some work has produced promising results on spontaneous speech [20, 21].

To avoid the ASR noise problems associated with free phone recognition and confidence measures, Extended Recognition Networks (ERN) generate a finite number of candidate errorful pronunciations using phonological rules and employs forced alignment to determine whether the candidate errorful or canonical pronunciations are more likely for each word [22, 23] or overall (for error tendency diagnosis) [24]. In [25], candidate pronunciations were automatically learned from the canonical pronunciation and spelling of each word, allowing end-to-end trainable error detection, though this was only evaluated on read speech. These methods use 1-best outputs so, unlike pronunciation scoring, are not probabilistic and so don't provide confidence estimates or allow incorporation of prior probabilities.

This paper builds on the above work by presenting a framework to explicitly divide pronunciation errors into accent and lexical errors, generate a dictionary of candidate pronunciations for each, perform lattice forced-alignment therewith and use pronunciation scoring on features from the resultant lattices for probabilistic error detection and tendency diagnosis of each type of error. This approach is investigated on three corpora.

2. Accent and lexical errors

Consider a speaker speaking a word w_i (e.g. the word *the*). Let the word's *intended pronunciation* be the sequence of phones $\boldsymbol{\psi}_{1:N}^{(w_i)} = \psi_1^{(w_i)}, \psi_2^{(w_i)} ... \psi_N^{(w_i)}$ that the speaker is trying to say (i.e. thinks is the correct lexical pronunciation of the word) and the *apparent pronunciation* be the sequence of phones $\boldsymbol{\phi}_{1:M}^{(w_i)} = \phi_1^{(w_i)}, \phi_2^{(w_i)} ... \phi_M^{(w_i)}$ that a listener would hear, given the way the speaker rendered the word.

The word w_i has a set of canonical pronunciations which a listener would recognise as correct, represented by the pronunciation dictionary entry $\mathcal{D}_{w_i}^{(0)}$ (e.g. $\mathcal{D}_{the}^{(0)}$ ={/dh ax/, /dh iy/}). A *lexical error* occurs when the intended pronunciation is not one of the canonical pronunciations (i.e. the speaker does not know the correct pronunciation of the word):

$$\boldsymbol{\psi}_{1:N}^{(w_i)} \notin \mathcal{D}_{w_i}^{(0)} \tag{1}$$

e.g. pronouncing *subtle* as /s ax b t l/.

An *accent error* occurs when the intended pronunciation does not match the apparent pronunciation (i.e. the speaker pronounces the correct phonemes in a way that sounds incorrect):

$$\boldsymbol{\phi}_{1:M}^{(w_i)} \neq \boldsymbol{\psi}_{1:N}^{(w_i)} \tag{2}$$

e.g. pronouncing *the* as /d ax/.

When processing speaker audio using an ASR, the intended pronunciation cannot be directly inferred, and it is only possible to ascertain the apparent pronunciation and compare it with the canonical pronunciations. However, there are other expected properties of the two types of errors that should make them distinguishable. Lexical errors depend on the graphemic rather than phonemic rendering of a word (e.g. someone who pronounces the silent b in subtle is not likely to also pronounce *scuttle* as /s k ax b t l/), while accent errors are uniquely a product of the canonical phone (e.g. someone pronouncing *the* as /d ax/ would be more likely to also pronounce *that* as /d ah t/). Separate models can thus be defined for the production of each, conditioned on canonical pronunciation for accent errors, and on spelling for lexical errors.

3. Proposed methodology

Consider an utterance with audio frames $\boldsymbol{o}_{1:T}$, which an ASR has recognised as containing the word sequence $w_{1:I}^{(ASR)}$, and in which we want to detect words pronounced incorrectly.

Starting with each possible canonical pronunciation $\phi_{1:M_j}^{(w_i)} \in \mathcal{D}_{w_i}^{(0)}$ of each word w_i, accent error candidates are generated by applying a sequence of up to R mispronunciations, namely discrete insertions, deletions and substitutions of phones, where possible. Specifically, these are: word-final deletion and insertion, consonant cluster reduction and anaptyxis [26], word-initial insertion, schwa lengthening, vowel shortening, dipthong reduction, consonant voicing, affricate confusion, affricate-fricative and fricative-plosive substitution, liquid consonant confusion and v-w and y-j substitution.

Lexical error candidates are generated by passing the spelling of w_i through a grapheme-to-phoneme (G2P) system trained on canonical pronunciations and removing canoncial pronunciations in the output. The hypothesis is that non-canonical pronunciations predicted by such a G2P are also likely to be made by a non-native speaker who hasn't encountered the word and guesses its pronunciation from its spelling.

The result is an entry $\mathcal{D}_{w_i}^{(R)}$, containing, as needed, accent errors of a particular type or all possible accent errors, lexical errors or both. The utterance $\boldsymbol{o}_{1:N}$ is now Viterbi forced aligned, replacing the canonical pronunciation dictionary for w_i with the union $\mathcal{D}_w^{(0)} \cup \mathcal{D}_w^{(R)}$. The alignment process generates a lattice, representing the likelihood $p(\boldsymbol{o}_{1:T}, \pi)$ of each possible path π through $\boldsymbol{o}_{1:T}$. Each path π corresponds to a sequence of phone labels $\phi_{1:M}^{(w_i)}$ (from the supplied dictionary) representing the apparent pronunciation of each word w_i and a sequence of start and end times $t_m^{(0)} : t_m^{(1)}$ for each phone ϕ_m.

The posterior probability that w_i was pronounced errorfully given it was recognised correctly by the ASR is given by:

$$p(e_n | \boldsymbol{o}_{1:T}, w_n^{(ASR)}) = \frac{\sum_{\pi | \phi_{1:M}^{(n)} \notin \mathcal{D}_{w_n}^{(0)}} p(\boldsymbol{o}_{1:T}, \pi)^{-\gamma}}{\sum_{\pi | w_n^{(\pi)} = w_n^{(ASR)}} p(\boldsymbol{o}_{1:T}, \pi)^{-\gamma}} \tag{3}$$

where γ is an acoustic model scaling factor.

Errors are detected by computing and thresholding this posterior at a value tuned for balance between precision and recall (in this work, this is achieved by maximising F1 score).

4. Corpora

Data Set	Trans.	#Utts.	#Words	#Marked errors
BLT	MAN	1438	61722	5968 (9.7%)
	CS	1438	53668	4546 (8.5%)
	ASR	1443	51535	4464 (8.7%)
SELL	MAN	149	3003	363 (12.1%)
	ASR	149	2701	296 (11.0%)
LeaP	MAN	45	6536	4982 (76.2%)
	ASR	43	6536	4383 (65.1%)

Table 1: *Annotated errors in each dataset, using original manual (MAN), ASR and crowd-sourced (CS) [27] transcriptions.*

Three corpora are investigated. The first consists of candidate recordings from the Business Language Testing Service (BULATS) spoken English test for foreign learners [28]. It comprises spontaneous speech, manually transcribed [29], with pronunciation errors and corrections annotated using ARPABET [30]. The dataset used in this work contains 226 speakers of varying proficiency, balanced for gender and between 6 L1s (Arabic, Dutch, French, Polish, Vietnamese and Thai).

Next, the SELL-CORPUS [31] consists of recordings of 389 volunteer Chinese speakers of English of varying proficiency, gender balanced and spread across 8 L1s/dialects (Northern Mandarin, Southwest Mandarin, Wu, Cantonese, Xiang, Minnan, Hakka and Gan), reading phonetically balanced utterances sampled from Project Guttenberg, with pronunciation errors and corrections human annotated using ARPABET.

Finally, the spontaneous part of the Learning Prosody in a Foreign Language (LeaP) project [32] is used. 50 non-native speakers of English, with 16 L1s, were recorded being interviewed before and after a prosody training course. It is phonetically annotated by humans using the X-SAMPA alphabet [33], unlike BULATS and SELL where annotators marked pronunciation errors.

For work on SELL, for which annotators assumed US pronunciations, the CMU [34] pronunciation dictionary was used. For BULATS and LeaP, a COMBILEX dictionary [35] of RP English was used instead.

5. Experimental Setup

To be able to detect errors in spontaneous speech, the first step is recognising the text being spoken and aligning the audio to a sequence of phones. Both 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 L1, the accuracy of standard commercial "off-the-shelf" ASR systems is too low for non-native learner English. Instead, an ASR system trained on non-native learners of English is used [36, 37].

The LeaP data was further pre-processed by performing speaker diarisation to remove the sections of interviewer speech, converting the X-SAMPA annotations to ARPABET, identifying annotated errors by looking up annotated pronunciations in the pronunciation dictionary and selecting corresponding corrections by minimising Levenshtein distance to the annotated errorful pronunciation.

6. Results and Discussion

Experiments are conducted to investigate whether the distinction between accent and lexical errors introduced in Section 2 is consistent with patterns in the data and evaluate the effectiveness of the method proposed in Section 3 to identify each.

6.1. Distinctness of accent and lexical errors

After speech recognition has been performed on each dataset, the canonical pronunciations of each recognised word are looked up in the dataset's canonical dictionary and a corresponding accent error dictionary generated as described in Section 3. The maximum number of mispronunciations per word R is chosen to satisfy computational complexity constraints ($R = 2$ here). Annotated errorful pronunciations in the dataset which appear in this dictionary are identified as accent errors.

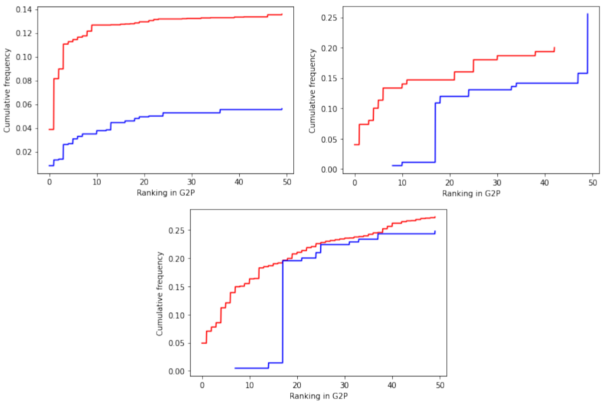

Figure 1: *Cumulative frequency of the ranking of identified accent errors (blue) and all remaining errors (red) present in BULATS (top, left), SELL (top, right) and LeaP (bottom) among the 50 first outputs of a G2P system trained on the respective canonical pronunciation dictionary of each corpus.*

Meanwhile, a Sequitur [38] G2P system, with a context window size of 3, is trained on the full canonical dictionary for each dataset and evaluated on each word, to produce a ranking of the 50 most likely pronunciations given each word's spelling. Each annotated error is looked up in this ranking and the cumulative frequency of rankings for each of the accent and, as yet, unidentified errors plotted (Figure 1). A ranking closer to 1 indicates that a pronunciation is more likely given the word's spelling.

It is seen that most accent errors are absent from the G2P output and that unidentified errors are more likely to rank better in the output than identified accent errors. This is consistent with the hypothesis that a significant fraction of the unidentified errors constitute lexical errors.

The experiment is repeated, splitting accent errors by type (Figure 2). It is seen that, across datasets, final deletions are the most likely to be predicted by G2P systems and voicing errors the least likely. However, as expected, all mispronunciation types rank below unidentified errors.

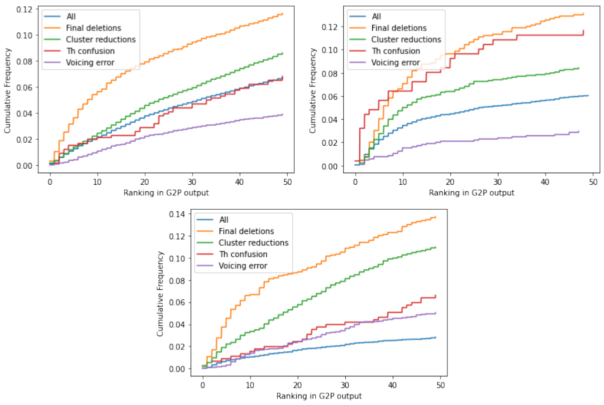

Figure 2: *Cumulative frequency of the ranking of a representative sample of types of identified accent errors in BULATS (top, left), SELL (top, right) and LeaP (bottom) among the 50 first outputs of the G2P system.*

A lexical error dictionary is now generated, as described in Section 3, keeping the first 10 pronunciations in each G2P output. The pronunciations in the accent and lexical error dictionaries are compared (Figure 3) and it is confirmed that the overlap is minimal. The dictionaries are then used to identiy errors among the annotated pronunciations (Figure 4).

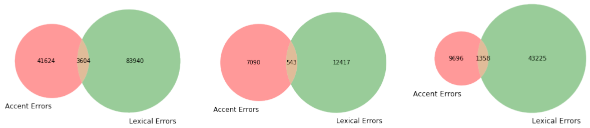

Figure 3: *Overlap of accent and lexical error candidate pronunciations in the dictionaries generated for the words in BULATS (left), SELL (middle) and LeaP (right).*

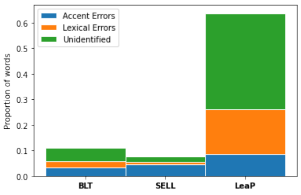

Figure 4: *Proportions of total words in each dataset annotated as errors and identified as accent and lexical errors.*

In BULATS and SELL, where annotators were asked to find and label pronunciation errors, only a small proportion of words are marked as errorful, more errors are identified as accent than lexical. By contrast, in LeaP, where annotators were asked to label the utterances phone-by-phone, the majority of words are marked with non-canonical pronunciations and most identified errors are lexical.

6.2. Detection performance

The accent and lexical error dictionaries generated for the words in the ASR outputs are now used to detect accent errors, specific types of accent errors, and lexical errors, as described in Section 3 (Figure 5, left). In addition to detection of errors at a word-by-word level, the sum of word-level posteriors (expected number of errors) is thresholded to detect the presence of each type of error at the utterance level (Figure 5, right).

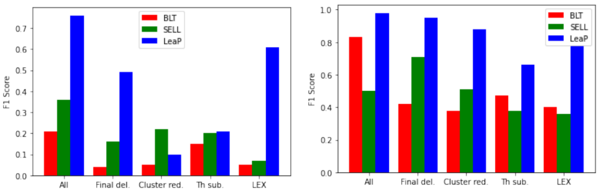

Figure 5: *F1 scores on each dataset for the task of detecting all accent errors ('All'), specific types of accent errors and lexical errors at a word-by-word (left) and utterance (right) level.*

It is seen that the system can accurately predict the presence of accent and lexical errors at the word-level for LeaP, but not for SELL and BULATS. At the utterance level, on the other hand, the system can diagnose tendency for accent errors, lexical errors and specific types of accent errors, across all three corpora.

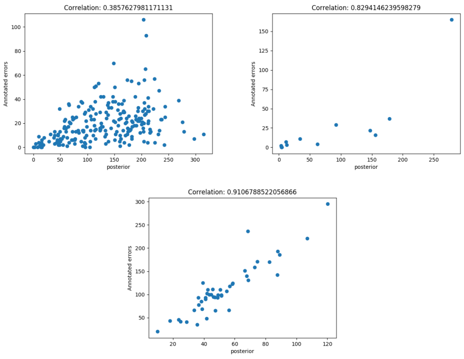

Figure 6: *Detected expected number of errors (sum of word-level posteriors) against annotated errors for each speaker in BULATS (top-left), SELL (top-right) and LeaP (bottom).*

Figure 6 shows the expected number of accent errors plotted against the number of actual annotated errors for each dataset. It is seen that the two correlate strongly, especially for SELL and LeaP and, in the cases of BULATS and SELL, more strongly than would be expected given the F1 scores of the prediction tasks. It is also noted that, for BULATS and SELL, the expected number of accent errors is much greater (almost an order of magnitude) than the actual annotated number of errors, while for LeaP this is not the case. All the above is consistent with the annotators in BULATS and SELL, who were instructed to specifically identify pronunciation errors, having annotated only a fraction of the accent errors actually present in the dataset.

This would explain why both utterance-level performance and correlation between aggregated word-level posteriors and number of annotated errors are high, while word-level performance is low. It would also explain why for LeaP, where annotators were instructed to label every single phone and the expected and annotated numbers of errors match, all three results are instead consistent and high.

6.3. Robustness to ASR error

Figure 7 shows the effect on the receiver operator characteristic (ROC) curves for accent error detection when the provided manual transcription is used instead of the ASR output. As expected, using the manual transcription yields an improvement, but the improvement is minor in all three cases. This is consistent with the system displaying robustness to ASR error.

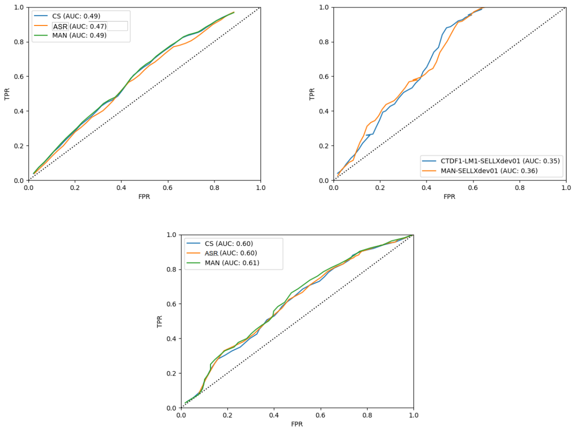

Figure 7: *ROC for the task of detecting accent errors in each of BULATS (top, left), SELL (top, right) and LeaP (bottom) using the ASR output, the provided manual transcription (MAN) and crowd-sourced transcriptions (CS).*

7. Conclusions

A framework for considering pronunciation errors as divided into accent and lexical errors and a methodology for detecting each is presented and evaluated. The framework is investigated in the context of three corpora, two on which humans were asked to annotate pronunciation errors, and one where they were asked to transcribe actual pronunciation. Results are consistent with accent and lexical errors being defined as distinct categories of error that can be detected separately. The system was successfully able to detect word-level accent and lexical errors on the latter corpus but not the former two. It was, however, able to diagnose lexical and general and specific accent error tendency with satisfactory performance across all three datasets. Analysis suggested that the annotators of the first two corpora were systematically under-annotating accent errors and that therefore the phonetic transcription technique is a superior method of annotation for error detection tasks.

8. Acknowledgements

This paper reports on research funded by Cambridge Assessment (CA), University of Cambridge. Thanks to CA for supporting this research and providing access to the BULATS data.

9. References

[1] Y. Xin, “The design of automatic correction system for spoken english pronunciation errors based on phonemes,” in *2019 6th Int. Conference on Dependable Systems and Their Applications (DSA)*. IEEE, 2020, pp. 247–254.

[2] Y. Feng, G. Fu, Q. Chen, and K. Chen, “SED-MDD: Towards sentence dependent end-to-end mispronunciation detection and diagnosis,” in *Proc. Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2020, pp. 3492–3496.

[3] M. Richardson, J. Bilmes, and C. Diorio, “Hidden-articulator Markov models for speech recognition,” *Speech Communication*, vol. 41, no. 2-3, pp. 511–529, 2003.

[4] J. Tepperman and S. Narayanan, “Using articulatory representations to detect segmental errors in nonnative pronunciation,” IEEE *Transactions on Audio, Speech, and Language Processing*, vol. 16, no. 1, pp. 8–22, 2007.

[5] L. Wei, W. Dong, B. Lin, and J. Zhang, “Multi-task based mispronunciation detection of children speech using multi-lingual information,” in *2019 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2019, pp. 1791–1794.

[6] R. Duan, T. Kawahara, M. Dantsuji, and H. Nanjo, “Cross-lingual transfer learning of non-native acoustic modeling for pronunciation error detection and diagnosis,” IEEE/ACM *Transactions on Audio, Speech, and Language Processing*, vol. 28, pp. 391–401, 2019.

[7] A. Diment, E. Fagerlund, A. Benfield, and T. Virtanen, “Detection of typical pronunciation errors in non-native English speech using convolutional recurrent neural networks,” in *Proc. Int. Joint Conference on Neural Networks (IJCNN)*, 2019, pp. 1–8.

[8] L. Zhang, Z. Zhao, C. Ma, L. Shan, H. Sun, L. Jiang, S. Deng, and C. Gao, “End-to-end automatic pronunciation error detection based on improved hybrid CTC/Attention architecture,” *Sensors*, vol. 20, no. 7, p. 1809, 2020.

[9] S. Robertson, C. Munteanu, and G. Penn, “Pronunciation error detection for new language learners.” in *Proc. INTERSPEECH*, 2016, pp. 2691–2695.

[10] Y. Kim, H. Franco, and L. Neumeyer, “Automatic pronunciation scoring of specific phone segments for language instruction,” in *Proc. Eurospeech*, 1997.

[11] O. Engwall, “Pronunciation analysis by acoustic-to-articulatory feature inversion,” in *Proc. Int. Symposium on Auto. Detection of Errors in Pronunciation Training*, 2012, p. 79.

[12] T. Nitta, S. Manosavan, Y. Iribe, K. Katsurada, R. Hayashi, and C. Zhu, “Pronunciation training by extracting articulatory movement from speech,” in *Int. Symposium on Automatic Detection of Errors in Pronunciation Training*, 2012, p. 75.

[13] K. Kamimura and K. Takano, “Pronunciation error detection in voice input for correct word suggestion,” in *Proc. of 2019 Int. Electronics Symposium (IES)*, 2019, pp. 490–493.

[14] A. Lee, N. F. Chen, and J. Glass, “Personalized mispronunciation detection and diagnosis based on unsupervised error pattern discovery,” in *Proc. Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2016, pp. 6145–6149.

[15] J.-C. Chen, J.-S. R. Jang, J.-Y. Li, and M.-C. Wu, “Automatic pronunciation assessment for Mandarin Chinese,” in *Proc. of the 2004 IEEE Int. Conference on Multimedia and Expo (ICME)*, vol. 3, 2004, pp. 1979–1982.

[16] S. Abdou, M. Rashwan, H. Al-Barhamtoshy, K. Jambi, and W. Al-Jedaibi, “Enhancing the confidence measure for an Arabic pronunciation verification system,” in *Proc. Int. Symposium on Auto. Detection of Errors in Pronunciation Training*, 2012, pp. 6–8.

[17] M. Nicolao, A. V. Beeston, and T. Hain, “Automatic assessment of English learner pronunciation using discriminative classifiers,” in *Proc. Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 5351–5355.

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

[19] H. Franco, H. Bratt, R. Rossier, V. Rao Gadde, E. Shriberg, V. Abrash, and K. Precoda, “Eduspeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications,” *Language Testing*, vol. 27, no. 3, pp. 401–418, 2010.

[20] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, “Automatic scoring of pronunciation quality,” *Speech communication*, vol. 30, no. 2-3, pp. 83–93, 2000.

[21] 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.

[22] M. Kane, Z. Ahmed, and J. Carson-Berndsen, “Underspecification in pronunciation variation,” in *Proc. Int. Symposium on Auto. Detection of Errors in Pronunciation Training*, 2012, p. 101.

[23] A. M. Harrison, W.-K. Lo, X.-j. Qian, and H. Meng, “Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training,” in *Proc. ISCA Workshop on Speech and Language Technology for Education (SLaTE)*, 2009.

[24] J. Lin, Y. Gao, W. Zhang, L. Wei, Y. Xie, and J. Zhang, “Improving pronunciation erroneous tendency detection with multimodel soft targets,” *Journal of Signal Processing Systems*, pp. 1–11, 2020.

[25] K. Li, X. Qian, and H. Meng, “Mispronunciation detection and diagnosis in L2 English speech using multidistribution deep neural networks,” IEEE/ACM *Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 1, pp. 193–207, 2017.

[26] G. Pelton, “Mining pronunciation data for consonant cluster problems,” in *Proc. Int. Symposium on Auto. Detection of Errors in Pronunciation Training*, 2012, p. 31.

[27] R. C. van Dalen, K. M. Knill, P. Tsiakoulis, and M. J. F. Gales, “Improving multiple-crowd-sourced transcriptions using a speech recogniser,” in *Proc. Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2015, pp. 4709–4713.

[28] L. Chambers and K. Ingham, “The BULATS online speaking test,” *Research Notes*, vol. 43, pp. 21–25, 2011.

[29] A. Caines, D. Nicholls, and P. Buttery, “Annotating errors and disfluencies in transcriptions of speech,” https://www.cl.cam.ac.uk/techreports/UCAM-CL-TR-915.pdf, University of Cambridge Computer Laboratory, Tech. Rep. UCAM-CL-TR-915, Dec 2017.

[30] A. Klautau, “ARPABET and the TIMIT alphabet,” 2001.

[31] Y. Chen, J. Hu, and X. Zhang, “Sell-corpus: an open source multiple accented Chinese-English speech corpus for L2 English learning assessment,” in *Proc. Int. Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, 2019, pp. 7425–7429.

[32] U. Gut, “The LeaP corpus: A multilingual corpus of spoken,” *Multilingual corpora and multilingual corpus analysis*, vol. 14, pp. 3–23, 2012.

[33] J. C. Wells, “Computer-coding the IPA: a proposed extension of SAMPA,” *Revised draft*, vol. 4, no. 28, p. 1995, 1995.

[34] R. Weide, “CMU pronunciation dictionary, rel. 0.6,” 1998.

[35] K. Richmond, R. Clark, and S. Fitt, “On generating Combilex pronunciations via morphological analysis,” in *Proc. INTERSPEECH*, 2010, pp. 1974–1977.

[36] K. Kyriakopoulos, M. Gales, and K. Knill, “Automatic characterisation of the pronunciation of non-native English speakers using phone distance features,” in *Proc. ISCA Workshop on Speech and Language Technology for Education (SLaTE)*, 2017.

[37] Y. Lu, M. J. F. Gales, K. M. Knill, P. P. Manakul, L. Wang, and Y. Wang, “Impact of ASR performance on spoken grammatical error detection,” in *Proc. INTERSPEECH*, 2019, pp. 1876–1880.

[38] C. G. Nevill-Manning and I. H. Witten, “Identifying hierarchical structure in sequences: A linear-time algorithm,” *Journal of Artificial Intelligence Research*, vol. 7, pp. 67–82, 1997.