Computer-Assisted Pronunciation Training: From Pronunciation Scoring Towards Spoken Language Learning

Nancy F. Chen* and Haizhou Li*†
* Institute for Infocomm Research, A*STAR, Singapore
E-mail: {nfychen,hli}@i2r.a-star.edu.sg
† National University of Singapore
E-mail:eleliha@nus.edu.sg

***Abstract*—This paper reviews the research approaches used in computer-assisted pronunciation training (CAPT), addresses the existing challenges, and discusses emerging trends and opportunities. To complement existing work, our analysis places more emphasis on pronunciation teaching and learning (as opposed to pronunciation assessment), prosodic error detection (as opposed to phonetic error detection), and research work from the past five years given the recent rapid development in spoken language technology.**

***Index Terms*—computer assisted language learning (CALL), second language learning, pronunciation tutoring, speech and language technologies in education**

I. Introduction

There is a re-emergence in computer-assisted pronunciation training (CAPT) research with the growing number of tech-savvy language learners, along with the tide of globalization, where many students are motivated by the increased economic opportunities of acquiring foreign languages. There has been several papers summarizing the research and develop efforts in using spoken language technology in education [1], [2], [3]. Eskenazi discusses a pronunciation tutoring prototype developed at CMU and the corresponding design considerations [1], where as [2] gives an overview of the field with a focus on modality interaction and a specific application for children. [3] is a more recent summary of automatic error detection in pronunciation training, which also includes a review of CAPT commercial systems. In this paper, we discuss ongoing research challenges, which focus more on pronunciation teaching and learning (as opposed to pronunciation assessment) and prosodic error detection (as opposed to phonetic error detection). We also summarize research work from the past five years given the recent rapid development in spoken language technology due to deep learning (e.g., [4], [5], [6], [7]).

Applications of CAPT can be divided into two areas: (1) pronunciation assessment, where the users are language evaluators, education centers, standardized testing services; (2) pronunciation learning/teaching, where the users are students and teachers. Pronunciation assessment has received some commercial success in bridging the gap between research and product development where the underlying technology is an acoustic model trained on native data of the target language, which are typically resource-rich languages like English. The goal of pronunciation assessment is to automatically score a non-native speech segment and produce results equivalent to a human teacher. Explicit detection of pronunciation errors is not essential.

On the other hand, developing automatic systems for pronunciation learning and teaching require much more linguistic resources. In addition to native data in the target language, non-native speech data annotated at the segmental (e.g., phonetic), subsegmental (e.g., place of articulation, manner of speech), or suprasegmental (prosodic) levels is essential for automatically detecting and characterizing the pronunciation errors. Most research and commercial development have focused on phonetic errors, but prosodic errors are often more detrimental to fluency perception [1]. The challenges in this domain range from multilingual acoustic modeling with limited resources to system integration of error detection types and multimodal user feedback [2], [3]. In this paper, we focus on discussing research challenges and opportunities for developing automatic systems that help students learn how to acquire spoken language skills.

II. Characterizing Pronunciation Patterns of Non-Native Speakers

Pronunciation errors are usually characterized at the phonetic (segmental) or prosodic (suprasegmental) level. Errors usually imply there is a clear-cut distinction between correct and incorrect pronunciation. The different speaking style of a non-native speaker is often referred to as having an accent, though one might still describe a native speaker's speech as having an accent when compared to a reference accent/dialect. Accents usually imply that the pronunciation difference is more fine-grained or harder to pinpoint (at the phonetic level).

A. Phonetic Errors

Phonetic errors are most commonly categorized as substitutions, insertions, or deletions. Deletions impact intelligibility the most, since more information is lost compared to substitutions and insertions.

Different phontactic constraints across languages might result in deletion and insertion errors. For example, in Vietnamese, only certain consonants are allowed at syllable final positions, so *face* might be pronounced as *fay* [8], [9]. Consonant clusters are not allowed in Vietnamese either, so vowels might be inserted in between consonants when Vietnamese speakers learn English.

According to language transfer theory, phonetic substitutions occur because of approximating L2 phonemes with L1 phonemes. In Mandarin and Spanish, there are no short vowels, so words like *eat* and *it* might sound similar when Mandarin and Spanish speakers speak English. Sometimes the non-native phone is neither in L1 or L2. It could be in between, or it could be very different. [10]. In these cases, it might make more sense to characterize the errors using subsegmental information according to articulatory gestures such as tongue position, lip configuration, manner of speech, voicing, and place of articulation.

Oftentimes for intermediate or advanced L2 learners, the phonetic errors are much more fine-grained than phonetic transformations. In [3], these phonetic errors are referred to as distortion errors. These distortion errors can contribute to the differences in the perceived accents.

B. Prosodic Errors

In terms of intelligibility, prosody is as important as phonetic accuracy, if not more important. A person with good segmental phonology who lacks correct timing and pitch is hard to understand, as intonation is the glue that holds a message together [1]. It indicates which words are important, disambiguates parts of sentences, and enhances the meaning with style and emotion.

Speech prosody is often characterized into three aspects: stress, rhythm, and intonation, though they are inextricably linked. Below we define the terminology and give examples of the possible prosodic errors in learning a second language.

1) Stress: Stress is the specific emphasis given to a particular syllable or word. Acoustically, speaking, a syllable with stress implies greater loudness, higher pitch, and longer duration. The stress placed on syllables within words are called *lexical stress* or *word stress*, while stress placed on words within sentences are called *sentence stress* or *prosodic stress*. Lexical stress is one of the most commonly investigated prosodic topics in automatic processing, since it is more well-defined.

Lexical stress placement in Bengali is *fixed* (restricted to the initial syllable of a word) [11], whereas English has *variable stress* (it is nontrivial to predict which syllables to place stress on). These differences in stress across languages can explain why Bengali speakers might place different lexical stress patterns than American English Speakers.

2) Rhythm: Rhythm refers to the temporal pattern of how a language is spoken. Languages such as English and German are *stress-timed*, which means some syllables are long while others (unstressed syllables) are short. For stress-timed languages, stress tends to occur at regular intervals, resulting in unstressed syllables being squeezed in between stressed syllables to accommodate the regular beat of the stress.

On the other hand, languages such as French and Spanish are *syllable-based*, meaning that each syllable is spoken at a regular interval. Therefore, when a French speaker learns English, part of his accent is due to the differences in rhythm of his native and foreign language.

3) Intonation: Intonation refers to the variation in pitch. Intonation helps the listener parse the boundaries in speech, suggesting when the sentence will end and whether contrastive information will be said next. Intonation also helps convey the speaker's attitude and emotions, such as uncertainty in questions or surprise in exclamations.

For tonal languages such as Mandarin Chinese and Vietnamese, variation in pitch can result in words with different meanings. This type of intonation is thus termed as *lexical tones*. L2 learners of Mandarin usually find the acquisition of lexical tones the most challenging, especially if the learner's first language is not tonal [12], [13], [14].

III. Research Approaches

A. Frameworks for Detecting Phonetic Errors

ASR is often a natural component in a CAPT system. An ASR system is used to provide confidence scores for error detection (e.g., likelihood-based scoring) or to diagnose the mispronounced error by forced-alignment with a lexicon that includes expected pronunciation patterns in the lexicon (extended recognition network). Even for systems that are not based on acoustic models derived from ASR [15], an ASR system is usually at least used to perform forced alignments to obtain phonetic boundaries of the non-native speech, since virtually all CAPT systems assume text-dependence.

The ASR system can be trained with just native L1 or with both non-native speech and native speech. The latter approach usually yields much better performance (e.g., 10% decrease in phone error rate [16]. In the very minimum, including the non-native data when training the acoustic model helps reduce the mismatches from channel or reading style.

In Table I, we compare four types of frameworks based on the following attributes: 1) If the approach's underlying model is ASR-based, 2) If the method requires L1 or L2 specific knowledge (i.e., how easy is it to port the approach to another language), 3) Is the method able to detect pronunciation errors (at the phonetic level), 4) If the method is able to diagnose what the mispronunciation is. Below, we elaborate on the approaches in more detail.

1) Likelihood-Based Scoring: Likelihood-based scoring approaches for phonetic error detection in computer-assisted language learning started in the 1990's [17], which showed that log posterior scoring scheme correlates the highest with

TABLE I
COMPARING DIFFERENT FRAMEWORKS FOR DETECTING PHONETIC ERRORS.

Framework	ASR-based	L1 Independence	L2 Independence	Error Detection	Error Diagnosis
Likelihood-based Scoring (GOP)	✓	✓	✓	✓	
Classifier-based Scoring		maybe	maybe	✓	✓
Extended Recognition Network (ERN)	✓			✓	✓
Unsupervised Error Discovery		✓	✓	✓	✓

human scores when compared to log-likelihood scores or segment duration scoring. Likelihood-based scoring is a defacto standard, and often referred to as *Goondness of Pronunciation (GOP)*, which was coined in [18], [19]. In practice, likelihood approaches are often used to quantify the difference between the likelihood score from forced alignment and the likelihood score from open phone loop decoding. The popularity of this approach stems from its L1 independence and ease to compute. The main drawback is that the likelihood score alone cannot provide diagnose what the mispronunciations are. The following approaches have thus been proposed to tackle this problem.

2) Classifier-Based Scoring: Classifier-based approaches typically target specific phoneme confusion pairs that represent common error types, so they are often dependent on knowledge of the L1 and/or L2. Truong et al used acoustic-phonetic features to train binary classifiers to distinguish confusion pairs from L2-learners of Dutch. The confusion pairs are extracted from the alignment between human transcribed non-native speech (surface pronunciation) and the canonical lexicon (underlying pronunciation)[20]. Classifier approaches can also be used for verification purposes once there are hypothesized detections, which could be derived from likelihood-based scoring (e.g., [6]). [21] compared a number of different experimental setups for detecting a confusion pair, including using different features (e.g., acoustic phonetic, MFCC, GOP).

3) Extended Recognition Networks (ERN): Extended recognition networks (ERN) employ an ASR setup but enhances the lexicon with possible/expected pronunciation error patterns, which are obtained either from consulting with experts or from non-native speech transcriptions [22]. ERN's can be viewed as a supervised learning approach to characterize errors of L2 learners beyond just targeted confusion pairs as in the typical classifier-based setting.

4) Unsupervised Error Discovery: A major challenge in CAPT research is the scarcity of large-scale non-native speech data with human annotations (at the phonetic level). Unsupervised error discovery approaches have thus been proposed [23], [15], where error patterns could be inferred by clustering acoustic segments and quantifying the distances among clusters of phonetic units labeled with the canonical representation. These unsupervised approaches detach the language dependency constraints of ERNs or classifier-based approaches, although the error detection performance is expected to be suboptimal when compared to supervised approaches due to the intrinsic noisier process of unsupervised learning.

B. Strategies for Improving Phonetic Error Detection

1) Verification/Rescoring: In a detection task, after postulating a (ranked) list of hypothesize detections, it is customary to perform a rescoring or verification step to further improve detection performance. Calibrating the dynamic ranges of detection scores to increase the separation margin between correct and incorrect detections has shown effective gains in detection tasks like spoken keyword search [24] and speaker and language recognition [25], [26].

In CAPT, there is an inevitable acoustic channel mismatch caused by different recording conditions between the training data (mostly native speech if not all) and the test data (non-native speech). While adding non-native data into the training data can help mitigate this mismatch to some extent, one usually does not have a lot of non-native data to spare. [15] used a speaker-dependent template-based speech recognizer for rescoring mispronunciation detections.

The classifier-based framework can be combined with the three other frameworks and serve as a verification step in pronunciation assessment and scoring. In other words, the three other frameworks can be viewed as different approaches to extract features for the classifier framework. For example, in [6] and [5], a neural network classifier with GOP scores as inputs were used to verify the mispronunciation detections.

2) Deep learning: Since the core technology is often ported from acoustic modeling in automatic speech recognition, CAPT benefits from the recent rapid advancement due to the emergence of deep learning. Reference [4] is one of the earliest studies to employ a deep learning framework for CAPT. The authors proposed the hybrid DBM-HMM acoustic model for mispronunciation and diagnosis and showed up to 18% improvement in word pronunciation error rate when compared to a GMM-HMM baseline. In [5], the researchers improved phone mispronunciation detection by using deep neural networks to obtain GOP likelihood scores for senones, which were fed to a neural network based logistic regression classifier. In [27], deep belief network (DBN) posteriorgrams were as input features to a dynamic time warping comparison system to detect word-level mispronunciations. Convolutional neural networks were used in [7] to automatically extract features for an MLP classifier for mispronunciation error detection, where the CNN filters could help one interpret the error patterns more intuitively.

3) Articulatory or Acoustic Phonetic Knowledge: While mainstream ASR approaches have helped advance CAPT, knowledge from acoustic phonetics or articulatory gestures could also be complementary and useful. For example, acous-

tic landmarks are time points on a speech spectrogram where there is sudden signal change [28]. There is rich acoustic information near acoustic landmarks due to the articulatory movements of speech production. Designing models from a landmark perspective, could be complementary to standard mainstream modeling approaches in ASR that implicitly assumes information is evenly distributed in the speech signal.

[29], [30] have exploited acoustic landmark-based SVM classifiers for detecting possible English pronunciation errors made by Korean learners, and showed that they complement HMM confidence scores. [31] also used landmarks to train an SVM classifier for detection correct and incorrect pronunciations of Mandarin nasal codas. They showed that the landmark-based classifier complements a DNN-HMM system.

In addition to using acoustic landmarks, one can also exploit acoustic phonetic properties or distinctive features [32], such as the place of articulation, the manner of speech, voicing, aspiration, vowel height. Some might call these properties speech attributes [33]. [6] used speech attributes to help further refine a DNN-HMM acoustic model to improve mispronunciation detection, and provide more informative feedback to the learner [34]. In [35], pronunciation error tendencies in terms of articulatory gestures were annotated and modeled to help give learners more instructive feedback.

C. Detecting Prosodic Errors

1) Lexical Stress: Lexical stress is the most well-studied prosodic error, where a classifier approach is most commonly adopted. Most research compares the importance of individual features and/or performance of various classifiers. We summarize some recent work below.

[36] conducted an exhaustive study comparing a number of English lexical stress classifiers using posterior probabilities and a range of prosodic features on classifying unstressed vowels, primary stress on vowels, and secondary stress on vowels spoken by L1-English and L1-Japanese children.They showed that Gaussian mixture models perform the best compared to decision trees and neural networks. [37] studied how different combinations of prosodic features perform using a classification and regression tree (CART) on French learners speaking German. They found that duration and pitch estimates are the most important features, which correspond with what is reported in the linguistic literature [38], [39]. [40] showed that prosody features that are context-aware consistently obtain higher classification accuracy using support vector machines (SVM) on Taiwanese learners speaking English. [41] used a deep belief network for detecting lexical stress in English spoken by L1-Mandarin and L1-Cantonese speakers. They showed that the proposed deep belief network improves over Gaussian mixture models when using prosodic features and expected lexical stress as features.

2) Lexical Tones: For tonal languages, lexical tones can be viewed as equivalents to lexical stress, but there are a number of differences. From an acoustic perspective, lexical tones are primarily characterized by the pitch contour (e.g., Mandarin), sometimes the pitch height (e.g., Cantonese), and sometimes glottal articulatory gestures (e.g., Vietnamese). From a phonological perspective, lexical tones are usually assigned to each syllable, while lexical stress is only assigned to some. For lexical stress, primary stress, secondary stress, and unstress syllables imply a gradient relationship of higher intensity and longer duration to lower intensity and shorter duration. On the other hand, for lexical tones, the differences among the different lexical tones are usually dictated by pitch contour, pitch height, or globalization differences.

There are typically two approaches for lexical tone recognition: one-step and two-step approaches. In a two-step approach, syllable boundaries are first identified before further modeling, while this does not exist in a one-step approach. [42] showed that a CD-GMM-HMM tone-based ERN framework outperforms the state-of-the-art two-step approach using TRUES (Tone Recognition Using Extended Segments) that models both unvoiced and voiced regions [43].

The approaches above have mainly been applied to native speech. Less work have focused on detecting lexical tone mispronunciations. Few exceptions include [44] and [45]. [44] proposed to segment the F0 contour to tone nucleus and articulatory transition regions to help improve lexical tone recognition accuracy. [45] proposed goodness of tone (GOT), a confidence measure using tonal phone (phoneme) likelihood modeling, which is inspired by GOP [18], [19]. The GOT features were modeled by an SVM classifier, which outperformed a Token-FFV baseline [46], which trains a Gaussian mixture model using fundamental frequency variation features [47] and then a 5-gram language model is trained with the GMM-indexes obtained from GMM tokenization.

There has also been recent work suggesting that pitch-related features could be inferred from a DNN system trained by 40-dimension MFCC features (instead of the usual 13) [48], [49], resulting in the MFCC system outperforming the F0 system in recognizing native Mandrin lexical tones. More investigation is needed to disentangle such counter-intuitive yet interesting observations by exploring possible sources of features to infer lexical tone information and examining the interaction between features and classifiers, especially those who employ deep learning approaches.

D. Automatic Fluency Scoring

At the end of the day, every L2 speaker wants to sound fluent. However, few studies have examined how to automatically predict fluency scores rated by humans, partially because few suitable corpora exist to support such research, despite its importance. [50] found that rate of speech (# phonemes/total duration of speech, including pauses) correlates highest with perceptual fluency. In addition, the number of silent pauses and the rate of articulation (#phonemes/total duration of speech without pauses).

Since non-native training data is often limited for fluency

scoring[1], models such as subspace Gaussian mixture models that have compact parameter sharing mechanisms could be useful. [16] used a subspace Gaussian mixture model trained with both phonetic and prosodic features to predict human rated fluency scores on read utterances spoken by non-native learners of Mandarin.

In most automatic speech assessment systems, the typical classifier-based setup is to manually engineer front-end features and then feed them into a machine-learning scoring model, so the two steps are done separately. In [51], the authors attempted to jointly optimize the feature learning and model scoring steps using a bidirectional LSTM recurrent network.

IV. Challenges and Research Opportunities

A. Scarcity of Large-Scale Linguistic Resources

Reference [13] compiled a list of non-native corpora that are of sufficient size (at least 100 hr, at least 100 speakers, or at least 10,000 utterances). The majority of the target languages (L2) are English, few have phonetic level transcriptions, and even fewer have proficiency ratings (fluency scores).

1) Lack of Non-Native Speech Data: To bypass the lack of suitable linguistic corpora, some researchers have used native audio to hypothesize the non-native mispronunciations. For example [52] simulated substitution phonemic errors by artificially introducing them in a native corpus. While [53], [54] shows that this approach works properly if the simulated errors reflect errors that are actually made by L2 learners, phonemic substitutions do not account for all errors. This approach assumes that the non-native mispronunciations fit nicely in well-defined categories in the target language L2, which is often not the case as distortion errors that are not easy to categorize are common [14]. For prosodic errors, this bypass approach might not be as straightforward, because this could require imposing intonation patterns from a native Bengali speaker to a native English speaker but ignoring other nuances such as the interaction between phonetic and prosodic patterns, and interaction among stress, energy level, intonation, rhythm, and underlying spoken content.

2) Lack of Human Annotations: For CAPT, phonetic-level transcription is often desirable since pinpointing phonetic errors is a major goal in automatic pronunciation teaching and learning. Yet compared to word transcriptions, phonetic transcriptions require more cost, time, and labor (linguistic expertise). Prosody labeling and fluency scoring can be much more subjective and harder to achieve inter-rater agreement [14].

For approaches that model articulatory gestures/speech attributes, the articulatory gestures are often inferred from the phonetic boundaries. This is a shortcut that could lead to reasonable CAPT results [6], but also ignores distortion errors and the asynchrony of articulatory gestures.

[1]Since fluency scoring needs to be at least at the utterance level instead of phonetic or word level, the number of datapoints to train classifiers that predict fluency scores is even less than those designed for detecting phonetic errors.

The motivation of unsupervised error discovery is to bypass the lack of linguistic resources, but these approaches typically still underperform when compared to supervised learning such as ERN [15]. Even if unsupervised learning can perform on par with supervised learning, we still need well-annotated data to assess its effectiveness in research.

B. Common Modeling Assumptions

1) Text dependence: While there has been work on spontaneous speech (e.g.,[55], [56]), the majority of work in CAPT implicitly assumes text-dependence, which is partially due to the even higher cost of human annotation of datasets if a CAPT system is text-independent. However, advanced learners need text-independent tools that can also detect errors beyond pronunciation, such as grammar and word usage. Only few reported systems are text-independent. [57] is an example where the users can create their own speech or text input to the CALL system. Another example is the SpeechRater engine developed at ETS scores non-native spontaneous English using both speech and natural language processing algorithms [58], [59].

A text-independent CAPT system provides the research opportunities to integrate automatic speech recognition, spoken language understanding, accent detection and characterization, mispronunciation detection and diagnosis, and recommendation systems for higher linguistic skills such as grammar, word usage, or topic coherence. It should be noted that there is a sizable number of advanced English learners that could greatly benefit from such systems.

2) Mispronunciations are Categorical: A major challenge in transcribing at the phonetic and tonal level is that non-native pronunciations might frequently fall out of the native phonemic or lexical tone categories; [10] describes three different ways a non-native phone might remain uncategorized: predominantly similar to one L1 phoneme, in between two L1 phonemes, or not similar to any L1 phonemes. In our experience, we have found similar (un)categorization issues with L2 phonemes and lexical tones in Mandarin [14].

Including an accent/dialect recognition and characterization module could be useful especially for advanced learners, or even native speakers that want to acquire a different accent [60], [61].

C. Metrics for Evaluation

1) Is lower mispronunciation detection error rate always better?: Evaluation metrics for mispronunciation detection error include the trade-off curve between false acceptance rate and false rejection rate. If one views the problem as an information retrieval task, one might examine precision and recall. From the pronunciation assessment perspective, the lower the mispronunciation detection error, the better. However, for pronunciation teaching, there are more aspects such as user feedback to consider. In terms of the trade-off curve for mispronunciation detection error, usually one would err on the end of high false acceptance rate instead of high false rejection rate, since the later is too damaging to a

learner's morale. While lower mispronunciation detection error rate might not need to be too low to be useful to the learner, a lower diagnostic error rate is helpful in giving correct user feedback.

2) Lack of User Studies: Metrics such as mispronunciation detection error and diagnostic error are widely used because they are easy to quantify and interpret. From the perspective of pronunciation teaching and learning, user studies are more directly helpful in evaluating the CAPT system, but so far very limited work has examined this aspect [57], [62].

How to prioritize which kind of feedback to provide, and how to provide it is not commonly studied. One might focus on mispronunciations that affect intelligibility the most if communication is the goal; one might focus on mispronunciations that affect fluency the most if the goal is to sound native; or one might focus on mispronunciations that are easier to correct if the learner feels frustrated. To improve CAPT systems for pronunciation learning and teaching, one can consider integrating research in human computer interface and audio/visual integration of mispronunciation detection and diagnosis.

V. Conclusions

This paper reviewed research approaches used in computer-assisted pronunciation training from both phonetic and prosodic perspectives. We also addressed existing research challenges related to the scarcity of linguistic resources, common modeling assumptions, and the lack of integration from human computer interface with spoken language technology. While automatic pronunciation assessment has reached a certain level of commercial success, there is still much potential for research and development opportunities in automatic pronunciation learning and teaching.

References

[1] Maxine Eskenazi, "Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype," *Language learning & technology*, vol. 2, no. 2, pp. 62–76, 1999.

[2] Maxine Eskenazi, "An overview of spoken language technology for education," *Speech Communication*, vol. 51, no. 10, pp. 832–844, 2009.

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

[4] Xiaojun Qian, Helen M Meng, and Frank K Soong, "The use of dbn-hmms for mispronunciation detection and diagnosis in l2 english to support computer-aided pronunciation training.," in *INTERSPEECH*, 2012, pp. 775–778.

[5] Wenping Hu, Yao Qian, Frank K Soong, and Yong 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.

[6] Wei Li, Sabato Marco Siniscalchi, Nancy F Chen, and Chin-Hui Lee, "Improving Non-Native Mispronunciation Detection and Enriching Diagnostic Feedback with DNN-Based Speech Attribute Modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2016, pp. 6135–6139.

[7] Ann Lee, *Language-Independent Models for Computer-Assisted Pronunciation Training*, Ph.D. thesis, MIT, 2016.

[8] Hoang Gia Ngo, Nancy F Chen, Binh Minh Nguyen, Bin Ma, and Haizhou Li, "Phonology-augmented statistical transliteration for low-resource languages," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[9] Hoang Gia Ngo, Nancy F Chen, Sunil Sivadas, Bin Ma, and Haizhou Li, "A minimal-resource transliteration framework for vietnamese.," in *INTERSPEECH*, 2014, pp. 1410–1414.

[10] Mona M Faris, Catherine T Best, and Michael D Tyler, "An examination of the different ways that non-native phones may be perceptually assimilated as uncategorized," *The Journal of the Acoustical Society of America*, vol. 139, no. 1, pp. EL1–EL5, 2016.

[11] Bruce Hayes and Aditi Lahiri, "Bengali intonational phonology," *Natural Language & Linguistic Theory*, vol. 9, no. 1, pp. 47–96, 1991.

[12] Nancy F. Chen, Vivaek Shivakumar, Mahesh Harikumar, Bin Ma, and Haizhou Li, "Large-scale characterization of Mandarin pronunciation errors made by native speakers of European languages," in *Interspeech*, 2013, pp. 803–806.

[13] Nancy F Chen, Rong Tong, Darren Wee, Peixuan Lee, Bin Ma, and Haizhou Li, "iCALL Corpus: Mandarin Chinese Spoken by Non-Native Speakers of European Descent," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[14] Nancy F. Chen, Darren Wee, Rong Tong, Bin Ma, and Haizhou Li, "Large-Scale Characterization of Non-Native Mandarin Chinese Spoken by Speakers of European Origin: An Analysis on iCALL," *Speech Communication*, 2016.

[15] Ann Lee, Nancy F Chen, and James Glass, "Personalized mispronunciation detection and diagnosis based on unsupervised error pattern discovery," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6145–6149.

[16] Rong Tong, Boon Pang Lim, Nancy F Chen, Bin Ma, and Haizhou Li, "Subspace Gaussian Mixture Model for Computer-Assisted Language Learning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2014, pp. 5347–5351.

[17] Yoon Kim, Horacio Franco, and Leonardo Neumeyer, "Automatic pronunciation scoring of specific phone segments for language instruction.," in *Eurospeech*, 1997.

[18] Silke Maren Witt, *Use of Speech Recognition in Computer-assisted Language Learning*, Ph.D. thesis, Cambridge University, 1999.

[19] Silke M Witt and Steve J Young, "Phone-level pronunciation scoring and assessment for interactive language learning," *Speech communication*, vol. 30, no. 2, pp. 95–108, 2000.

[20] Khiet Truong, Ambra Neri, Catia Cucchiarini, and Helmer Strik, "Automatic pronunciation error detection: an acoustic-phonetic approach," in *InSTIL/ICALL Symposium 2004*, 2004.

[21] Helmer Strik, Khiet Truong, Febe De Wet, and Catia Cucchiarini, "Comparing different approaches for automatic pronunciation error detection," *Speech Communication*, vol. 51, no. 10, pp. 845–852, 2009.

[22] Alissa M Harrison, Wai-Kit Lo, Xiaojun Qian, and Helen Meng, "Implementation of an extended recognition network for mispronunciation detection and diagnosis in computer-assisted pronunciation training.," in *SLaTE*, 2009, pp. 45–48.

[23] Ann Lee and James Glass, "Mispronunciation detection without nonnative training data," in *Sixteenth Annual Conference of the International Speech Communication Association*, 2015.

[24] Nancy F Chen, Haihua Xu, Xiong Xiao, D Van Hai, Chongjia Ni, I-Fan Chen, Sunil Sivadas, Chin-Hui Lee, Eng Siong Chng, Bin Ma, et al., "Exemplar-inspired strategies for low-resource spoken keyword search in swahili," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 6040–6044.

[25] Tomi Kinnunen and Haizhou Li, "An overview of text-independent speaker recognition: From features to supervectors," *Speech communication*, vol. 52, no. 1, pp. 12–40, 2010.

[26] Haizhou Li, Bin Ma, and Kong Aik Lee, "Spoken language recognition: from fundamentals to practice," *Proceedings of the IEEE*, vol. 101, no. 5, pp. 1136–1159, 2013.

[27] Ann Lee, Yaodong Zhang, and James Glass, "Mispronunciation detection via dynamic time wrapping on deep belief network-based posteriorgrams," in *ICASSP*, 2013.

[28] Kenneth N Stevens, "Toward a model for lexical access based on acoustic landmarks and distinctive features," *The Journal of the Acoustical Society of America*, vol. 111, no. 4, pp. 1872–1891, 2002.

[29] Su-Youn Yoon, Mark Hasegawa-Johnson, and Richard Sproat, "Automated pronunciation scoring using confidence scoring and landmark-based svm.," in *Interspeech*, 2009, pp. 1903–1906.

[30] Su-Youn Yoon, Mark Hasegawa-Johnson, and Richard Sproat, "Landmark-based automated pronunciation error detection.," in *Interspeech*, 2010, pp. 614–617.

[31] Yanlu Xie, Mark Hasegawa-Johnson, Leyuan Qu, and Jinsong Zhang, "Landmark of mandarin nasal codas and its application in pronunciation error detection," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 5370–5374.

[32] Kenneth N Stevens, *Acoustic phonetics*, vol. 30, MIT press, 2000.

[33] Roman Jakobson, Gunnar Fant, and Morris Halle, "Preliminaries to speech analysis. the distinctive features and their correlates," 1951.

[34] Wei Li, Li Kehuang, Sabato Marco Siniscalchi, Nancy F Chen, and Chin-Hui Lee, "Detecting mispronunciations of l2 learners and providing corrective feedback using knowledge-guided and data-driven based decision trees," in *Interspeech*, 2016.

[35] Yingming Gao, Yanlu Xie, Wen Cao, and Jinsong Zhang, "A study on robust detection of pronunciation erroneous tendency based on deep neural network," in *16th Annual Conference of the International Speech Communication Association (INTERSPEECH)*, 2015, pp. 693–696.

[36] Luciana Ferrer, Harry Bratt, Colleen Richey, Horacio Franco, Victor Abrash, and Kristin Precoda, "Classification of lexical stress using spectral and prosodic features for computer-assisted language learning systems," *Speech Communication*, vol. 69, pp. 31–45, 2015.

[37] Anjana Sofia Vakil and Jürgen Trouvain, "Automatic classification of lexical stress errors for german capt," 2015.

[38] A Cutler, DB Pisoni, and RE Remez, *The handbook of speech perception*, Blackwell Oxford, UK, 2005.

[39] Grzegorz Dogil and Briony Williams, "5 the phonetic manifestation of word stress," *Word prosodic systems in the languages of Europe*, p. 273, 1999.

[40] Junhong Zhao, Hua Yuan, Jia Liu, and S Xia, "Automatic lexical stress detection using acoustic features for computer assisted language learning," *Proc. APSIPA ASC*, pp. 247–251, 2011.

[41] Kun Li, Xiaojun Qian, Shiyin Kang, and Helen Meng, "Lexical stress detection for l2 english speech using deep belief networks.," in *INTERSPEECH*, 2013, pp. 1811–1815.

[42] Changliang Liu, Fengpei Ge, Fuping Pan, Bin Dong, and Yonghong Yan, "A one-step tone recognition approach using msd-hmm for continuous speech.," in *INTERSPEECH*, 2009, pp. 3015–3018.

[43] Jiang-Chun Chen and Jyh-Shing Roger Jang, "Trues: Tone recognition using extended segments," *ACM Transactions on Asian Language Information Processing (TALIP)*, vol. 7, no. 3, pp. 10, 2008.

[44] Jinsong Zhang and Keikichi Hirose, "Tone nucleus modeling for chinese lexical tone recognition," *Speech Communication*, vol. 42, no. 3, pp. 447–466, 2004.

[45] Rong Tong, Nancy F. Chen, Bin Ma, and Haizhou Li, "Goodness of Tone (GOT) for Non-native Mandarin Tone Recognition," in *Interspeech*, 2015, pp. 801–804.

[46] Rong Tong, Nancy F Chen, Boon Pang Lim, Bin Ma, and Haizhou Li, "Tokenizing fundamental frequency variation for mandarin tone error detection," in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5361–5365.

[47] Kornel Laskowski, Jens Edlund, and Mattias Heldner, "An instantaneous vector representation of delta pitch for speaker-change prediction in conversation dialogue system," in *ICASSP*, 2008.

[48] Neville Ryant, Jiahong Yuan, and Mark Liberman, "Mandarin tone classification without pitch tracking," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2014, pp. 4868–4872.

[49] Neville Ryant, Malcolm Slaney, Mark Liberman, Elizabeth Shriberg, and Jiahong Yuan, "Highly accurate mandarin tone classification in the absence of pitch information," in *Proceedings of Speech Prosody*, 2014, vol. 7.

[50] Catia Cucchiarini, Helmer Strik, and Lou Boves, "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology," *The Journal of the Acoustical Society of America*, vol. 107, no. 2, pp. 989–999, 2000.

[51] Zhou Yu, Vikram Ramanarayanan, David Suendermann-Oeft, Xinhao Wang, Klaus Zechner, Lei Chen, Jidong Tao, Aliaksei Ivanou, and Yao Qian, "Using bidirectional lstm recurrent neural networks to learn high-level abstractions of sequential features for automated scoring of non-native spontaneous speech," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 338–345.

[52] Joost van Doremalen, Catia Cucchiarini, and Helmer Strik, "Automatic detection of vowel pronunciation errors using multiple information sources," in *Automatic Speech Recognition & Understanding, 2009. ASRU 2009. IEEE Workshop on*. IEEE, 2009, pp. 580–585.

[53] Catia Cucchiarini, Ambra Neri, and Helmer Strik, "Oral proficiency training in dutch l2: The contribution of asr-based corrective feedback," *Speech Communication*, vol. 51, no. 10, pp. 853–863, 2009.

[54] Sandra Kanters, Catia Cucchiarini, and Helmer Strik, "The goodness of pronunciation algorithm: a detailed performance study.," *SLaTE*, vol. 2009, pp. 2–5, 2009.

[55] Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[56] Derrick Higgins, Xiaoming Xi, Klaus Zechner, and David Williamson, "A three-stage approach to the automated scoring of spontaneous spoken responses," *Computer Speech & Language*, vol. 25, no. 2, pp. 282–306, 2011.

[57] Hongcui Wang, Christopher J Waple, and Tatsuya Kawahara, "Computer assisted language learning system based on dynamic question generation and error prediction for automatic speech recognition," *Speech Communication*, vol. 51, no. 10, pp. 995–1005, 2009.

[58] Klaus Zechner, Derrick Higgins, Xiaoming Xi, and David M Williamson, "Automatic scoring of non-native spontaneous speech in tests of spoken english," *Speech Communication*, vol. 51, no. 10, pp. 883–895, 2009.

[59] Automated Scoring of Speech, "http://www.ets.org/research/topics/as_nlp/speech," last accessed, September 27, 2016.

[60] Nancy F Chen, Sharon W Tam, Wade Shen, and Joseph P Campbell, "Characterizing phonetic transformations and acoustic differences across english dialects," *Audio, Speech, and Language Processing, IEEE/ACM Transactions on*, vol. 22, no. 1, pp. 110–124, 2014.

[61] Hamid Behravan, Ville Hautamäki, and Tomi Kinnunen, "Foreign accent detection from spoken finnish using i-vectors.," in *INTERSPEECH*, 2013.

[62] Yasushi Tsubota, Tatsuya Kawahara, and Masatake Dantsuji, "Practical use of english pronunciation system for japanese students in the call classroom," in *Proc. ICSLP*, 2004, vol. 15, pp. 1689–1692.