Automatic Error Detection in Pronunciation Training: Where we are and where we need to go

Conference Paper · June 2012

CITATIONS
62

READS
3,195

1 author:

Silke Maren Witt
Amazon Inc.
24 PUBLICATIONS **657** CITATIONS

SEE PROFILE

Automatic Error Detection in Pronunciation Training: Where we are and where we need to go

Silke M. Witt
Fluential, Inc
Sunnyvale, USA
switt@fluentialinc.com

***Abstract* — This paper discusses the state of the art of research in computer assisted pronunciation teaching as of early 2012. A discussion of all major components contributing to pronunciation assessment is presented. This is followed by a summary of existing research to date. Additionally, an overview is given on the use of this research in commercial language learning software. This is followed by a discussion of remaining challenges and possible directions of future research.**

Keywords – Pronunciation error detection, automated error detection, computer assisted language learning, Computer Assisted Pronunciation Training (CAPT)

1 Introduction

In the wake of tremendous improvements in computing power and multi-modal applications, there has also been a renewed interest in computer-assisted pronunciation teaching (CAPT) applications in recent years. With increasing globalization, there has also been a significant increase in the demand for foreign language learning, one aspect of which is pronunciation learning. Effectively teaching pronunciation typically requires one-to-one teacher student interactions, which for many students is unaffordable. For this reason, automatic pronunciation teaching has been a focus of the research community, bringing together researchers from a number of disciplines: speech recognition, linguistics, psycholinguistics, and pedagogy, as well as auditory and articulatory research.

2 A short History of Capt

Research work on automated pronunciation error detection and pronunciation assessment started in 1990's with a flurry of activities in late 90's to early 2000, see references [1] to [8] and [9]. A detailed list of early references can also be found in [10]. Since there are a large number of publications in this area, it was only possible to quote some representative examples of papers. The author apologizes for any omissions. In the early 2000's commercialization of CAPT proved difficult and thus research activities, too, slowed down. With increased computing power, mobile devices and improved speech recognition, interest picked up again about five years ago, leading to the founding of an ISCA special interest group called SLaTE (Speech and Language Technology for Education) in 2007.

References [11], [12] and [13], provide a very thorough and in-depth overview of the work up to 2009. Since pronunciation error detection and teaching in its entirety is a difficult problem, past work has often only addressed components of this field such as phoneme level pronunciation error detection or prosodic error detection.

The next section will discuss the different components that contribute to pronunciation, followed by a discussion of many features that have been proposed to measure these pronunciation components. Then section 4 presents a discussion of existing research for these various components, followed in section 5 by an overview of commercial systems that employ some of this research. Section 6 then discusses remaining challenges in CAPT.

3 What is pronunciation?

Pronunciation is a general term that covers a number of different components and can be measured with many different features. In addition, the term 'pronunciation error' is difficult to quantify; i.e. there is no clear definition of right or wrong in pronunciation. Rather there exists an entire scale ranging from unintelligible speech to native-sounding speech.

3.1 Types of Pronunciation Errors

Figure 1 below illustrates all different aspects of pronunciation errors that need to be addressed in a successful training and assessment situation. Pronunciation errors can be divided into phonemic and prosodic error types.

On the phonemic side there are the 'severe' errors where phonemes might be substituted with another phoneme, deleted or inserted. Then there are the 'errors' on a smaller scale where the correct phoneme is more or less being spoken, however, the sound of it is still different enough from a native speaker's pronunciation that it is noticeable that a speaker still has an accent.

On the prosodic side a non-native accent can be categorized in terms of stress, rhythm and intonation. All such errors are closely linked which is indicated by the circles in Figure 1. This fact makes pronunciation a multi-dimensional problem that is difficult to pin down with a single approach. Rather, a successful system will require a combination of many different techniques.

One additional challenge in pronunciation error detection is that a phoneme represents the smallest possible unit compared

to the syllable, word and sentence level. The shorter the unit, the higher will be the variability in the judgment of the pronunciation quality. Even human judges have been shown to have low inter-rater and intra-rater correlations. Consequently pronunciation error detection at the phoneme level is a much harder task than measuring pronunciation fluency across multiple sentences. Kim et al. [7] showed that the error detection accuracy can be significantly improved if all realizations of the same phoneme in a student's speech sample are scored at speaker level, that is the final phoneme score is an average of all same phoneme occurrences of one speaker. They found that after about 300 instances of a given phoneme, the correlation with human ratings reaches 0.8, while the correlation of the rating for a single phoneme instance is as low as 0.5. This approach however has the disadvantage of ignoring the fact that pronunciation errors depend on surrounding phonemes and spelling-to-pronunciation rules.

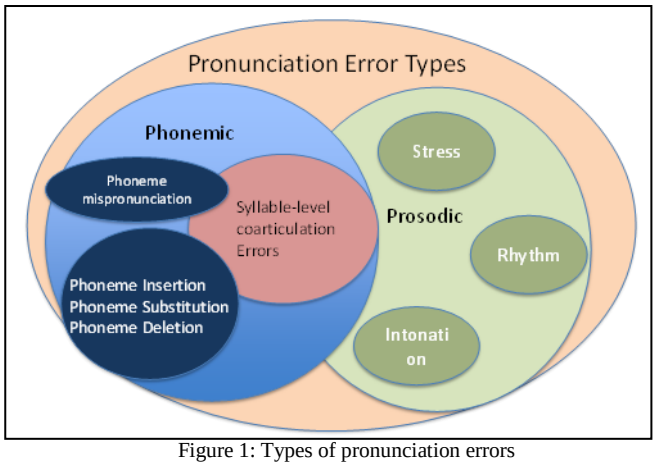

Figure 1: Types of pronunciation errors

3.2 *Native-like versus intelligible pronunciation*

In recent years, a discussion has started whether the goal of pronunciation teaching is to sound just like a native speaker or whether the teaching should mostly focus on the intelligibility of the student. Currently, the agreement appears to be that intelligibility is an essential component of communicative competence. While aiming to sound just like a native speaker is important, especially for more advanced students, see [14], it is clearly less critical than basic intelligibility. Raux et al. [15] explored the relationship between error rates and intelligibility and found that errors related to prosodic features, such as vowel insertion, impact intelligibility more than segmental errors, such as phoneme substitution, for example replacing 'A' with 'ER". The authors present a probabilistic model that helps to predict intelligibility based on the student's errors.

Koniaris et al. [16], used a model of the human auditory system to identify those pronunciation errors that are most noticeable to native speakers. This model provides a distance measure between non-native and native speakers which takes into account the perception of sounds by native speakers.

3.3 *Pronunciation features*

One possibility is to visualize the pronunciation of a unit, phoneme, syllable, word or sentence as a cluster in an N-dimensional space, where each dimension represents a pronunciation feature. The values for each pronunciation feature for native speakers will vary within a given range. Thus taking all features with their variable range together, native pronunciation can be seen as an irregularly-shaped cluster in the N-dimensional space. With this image in mind, the task of assessing non-native pronunciation can be seen as measuring the distance of the non-native speaker's pronunciation to the aforementioned cluster of native speakers. The smaller this distance, the more 'native-like' the pronunciation will be.

A large number of metrics for measuring these dimensions of pronunciation has been used over the years. An excerpt of such metrics is shown in Table 1 below. The table content is by no means complete; it is intended to demonstrate the large variety of metrics and thus the large number of pronunciation dimensions. Particularly on the prosodic side there are many similar, but slightly differently defined features that have been used.

Feature Category	Feature Name
Phonemic	Phone-level log-likelihood scores, GOP
	Vowel durations, duration trigrams
	Phoneme pair classifiers
	spectral features (formants)
	Articulatory-acoustic features
Prosodic (Intonation, Stress, fluency)	distances between stressed and unstressed syllables
	Mean, max, min power per word (energy)
	F0 contours (slope and maximum)
	rate of speech (words per second/minute)
	Trigram models to model phoneme duration in context
	Phonation/time ratio, mean phoneme duration
	Articulation Rate (phonemes/sec)
	Mean and standard deviation of long silence duration
	Silences per second
	Frequency of disfluencies (pauses, fillers etc)_
	Total and mean pause time (i.e. duration of interword pauses)

Table 1: Examples of features used for pronunciation scoring

The next section will discuss in detail different metrics for the types of pronunciation errors using these features in various ways.

4 EXISTING RESEARCH ON PRONUNCIATION ERROR DETECTION

4.1 *Likelihood-based scoring*

The initial work on this topic in the 1990's saw the creation of several likelihood-based phoneme-level error detection algorithms. For example, Kim et al. [7] presented three HMM-based scores: a) a HMM-based log-likelihood score, b) a HMM-based log posterior score, which later on has become a de-facto standard, since it was shown that it had the highest

correlation with human scores (this work scores the pronunciation quality of a given phoneme over many instances of the pronunciation of a given phoneme), and c) a third score based on segment duration. Similarly, the 'GOP' (goodness of pronunciation) score also uses a log-likelihood based score, [10]. Likewise Kawai et al. [17] also used log-likelihood scores in forced alignment mode. Expanded versions of likelihood-based scores were also successfully used by Mak et al. [18].

4.2 L1-independent approaches

One of the core decision points for pronunciation error detection is whether to build a system that is L1 (i.e. the native language) dependent or not. While it is preferable to have an L1 independent system in order to minimize the commercial implementation challenges, better performance has been found with methods that take L1 into account. In addition to likelihood-based scores, which are in most cases L1-independent, there are several additional scoring approaches that are L1-independent. Cucchiarini et al. [19] also utilized a manually annotated corpus of non-native speakers of Dutch to generate statistics on both the frequency and the context of pronunciations errors. They showed that there is a good amount of overlap between manually derived error types in the linguistic literature and such automatically derived error types. In a recent approach, Li et al. [64] combined log-likelihood scores and fluency scores, like rate of speech and trigram phone duration model, in order to score pronunciation and was able to achieve a correlation at sentence-level with human ratings of 0.84. Similarly, Cincarek et al. [20] uses a classifier-based approach that combines log-likelihood and different duration scores in order to calculate mispronunciation probabilities of phonemes across multiple utterances. Lastly, Cincarek et al. [20] also applied a L1-independent scoring mechanism based on a combination of loglikelihood and duration scores to identify common mispronunciation patterns for a given language.

4.3 L1 Dependency

There exists a fairly large number of work that is L1-dependent since that approach has traditionally yielded a higher accuracy than L1-independent approaches. For example, Ito et al. [21] manually derived a set of mispronunciation rules for a given L1/L2 pair and used those for clustering error rules using a decision tree. This approach resulted in an increased pronunciation error detection accuracy.

4.3.1 Automatic generation of data for L1/L2 pairs

Taking into account L1 has two main advantages: Firstly, if L1 is known, one can utilize acoustic models that are a mixture of L1 and L2 ([10], [22], [23]) and have improved speech recognition accuracy, which in turn enables recognition of less constrained utterances, which allows for greater freedom in the selection of pronunciation learning exercises, in particular for assessing fluency. Secondly, the set of common pronunciation errors tend to be typical for a given L1 and very different between different L1, i.e. a German speaker will make very different English pronunciation errors than a native speaker of Chinese or Hindi. Thus, knowledge of L1 enables to provide tailored pronunciation exercises. For example Husby et al. [24], created a tool called L1-L2*map* that contains manually entered data on likely mispronunciations for a given L1 when learning Norwegian. This data was then used to create a list of expected pronunciation errors. Likewise, Neri et al. [25] conducted a similar analysis to identify L1 specific groups of commons errors for students of Dutch.

In recent years, there have been a number of approaches to automate the process of identifying typical error patterns for a given L1-L2 pair. Lo et al. [26] and Harrison et al. [27, 28] have utilized an alignment of canonical pronunciations with manually annotated pronunciations of non-native speech to automatically generate mispronunciation rules. Such rules are then used in extended recognition networks to identify pronunciation errors. One advantage of using these recognition networks is that if an error is identified the type of error is also known and can be used for diagnosis.

Quian et al. [29] explored an alternative method to generate mispronunciation lexica. They used joint sequence multigrams to perform a grapheme to mispronunciation conversion and showed that this approach can slightly improve performance both in terms of accuracy as well as reduction in false alarm and false rejection rates. However, all these approaches still require a manually annotated corpus of non-native speech which is expensive and time-consuming to create.

Along the same lines, Stanley et al. [30] conducted research in finding mechanisms to automatically model phonological errors. The authors showed that applying statistical machine translation significantly improved the precision and recall for pronunciation errors, while the accuracy was similar to the accuracy of the extended recognition networks.

4.4 Classifier-based scoring

While likelihood-based pronunciation scoring has the advantage of being L1-independent and very easy to compute, it has been found that the calculated scoring is not capable of identifying the error type that has occurred.

In order to address this problem, there have been a number of studies that employ classifiers for specific phoneme pair contrasts that represent common error types. For example, Franco et al. [31] built a set of classifiers for Dutch vowel contrasts and found that adding MFCC as well as phonetic features in addition to ASR features to train classifiers gave the best classification results. Likewise, Truong et al. [32] developed a L1-independent classifier utilizing a number of acoustic-phonetic features for each expected phoneme error combination. This classifier has been shown to outperform previous approaches, but does have the drawback that common errors for a given L2 have to be known and that separate classifiers for each error type are necessary. A similar classifier for Norwegian is presented by Amdal et al. [33].

For more recent work on error detection with the help of classifiers, see Strik et al. [34]. The authors compared the scoring accuracy for four different classifiers for a set of manually identified problem phoneme pairs for non-native

speakers of Dutch. This work demonstrated that LDA based classifiers can outperform log-likelihood-based scoring.
Similarly, Yoon et al. [35] trained a landmark-based SVM classifier on an expected set of distortion errors, where a landmark is a sudden signal change such as a stop release. This approach, however, required knowledge of mispronunciation rules for a given L1 – L2 language pair.

4.5 Non-native acoustic modeling

If a CAPT system allows freely spoken utterances from the student, non-native acoustic modeling is required. Hui et al. [22] showed that using standard adaptation algorithms such as MAP or MLLR yields substantial recognition accuracy improvements.
Likewise, Saz et al. [23] showed that going from speaker independent to speaker dependent recognition via MAP almost reduces the phoneme recognition error rate in half. Interestingly, there was little difference if the adaptation was conducted on all available material for a given speaker or if words that had labeled mispronunciations were excluded. Such results are encouraging, because it shows that un-supervised adaptation, even if it adapts to acoustic data that includes pronunciation errors, still yields a significantly better recognition performance than no adaptation.

4.6 Text independence

Up to now, little work has been attempted to assess the pronunciation quality of unconstrained spontaneous speech. However, for more advanced pronunciation learning activities, it is a requirement to have students speak text freely as opposed to reading a text.
In order to do so it has been proposed to use a sequence of two different recognition tasks, see the work by Moustrofas et al. [36] and Chen et al. [37]. First, the non-native speech has to be recognized irrespective of any pronunciation errors. This is typically done with acoustic models adapted to the particular characteristics of the speaker. Secondly, the recognized text is used to perform recognition in forced-alignment mode and to calculate the pronunciation 'correctness' based on one of the many algorithms proposed for this task.

4.7 Prosodic pronunciation error detection and feedback

A very detailed discussion of CAPT systems that provide prosodic feedback can be found in [12].
Bernstein et al. [38] have shown that there appears to be a linear relationship between fluency measures (such as listed in Table 1) and human judgments of proficiency. Also human-ratings of fluency have been found to be reliable with inter-rater correlation above 0.9, [3]. These results show the importance of measuring fluency as part of any pronunciation assessment exercise. Recently, there has been more interest in exploring automated methods to measure prosodic features of pronunciation. For example, Levow et al. [39], used a SVM-based classifier for pitch accent recognition. Hönig et al. [65] used a large feature set based on duration, energy, pitch and pauses to detect word accents. In more recent work, Hönig et al. [66] employed a discriminative approach that uses a large number of specialized rhythm features as well as general prosodic features to create a comprehensive metric of prosodic pronunciation quality. Bonneau et al. [40] presented a system that teaches fluency with several different methods of modifying the phoneme durations and F0 contour of a learner's speech in order to demonstrate to the student (in their own voice) what their pronunciation should sound like. Initial results from a pilot study seem promising, but a larger follow-up study is needed to confirm the initial findings.

Another aspect of pronunciation not discussed so far, are tones in tonal languages like Chinese. Mixdorff et al. [41] and Hussein et al. [42] conducted initial work to detect tone errors by German learners of Mandarin Chinese. The main challenged encountered here (as in a number of other CAPT system) was a high rate of false hits.

Very recent work by Engwall [43] uses audiovisual articulatory feature inversion to estimate the learner's current articulation and to provide audiovisual feedback by showing the movement of the tongue with computer animations. An initial evaluation seemed to show that such feedback on tongue movements helped students to improve their tongue position and thus their articulation and pronunciation.

4.8 Corrective feedback

Corrective feedback can only be effective, if the student is also able to perceive the difference non-native and native speech, see [42, 24, 44]. For example, before being able to learn the tones of Mandarin Chinese, a student must be able to perceive the tones. Thus, a CAPT system needs to include perception training as part of corrective feedback components.
One of the earlier systems that not only attempts to detect mispronunciations but also give the student some information as to how to correct the mispronunciation, is the PLASER system, [18]. While students liked the system and 77% of the participants of a test study believed their pronunciation to have improved, robustness and correct error detection at a phoneme-level were identified as problem areas.
Bodnar et al. [45] tested the feasibility of ASR-based corrective feedback via a virtual teacher with regard to teaching L2 syntax and found to be effective within the scope of a small test study. The future goal for this system is to build it out as a platform to test a variety of different feedback strategies. Similarly, the Euronounce system by Demenko et al. [46] uses several methods of corrective feedback, but this time the feedback focusses on prosody with the help of 'pitch-line', an approximation of intonation contours that attempts to only show the relevant components of intonation (pitch-accents, boundary tones).
In order for corrective feedback around tongue positioning to work, Ouni [47] has shown that if students receive short, specific training for tongue gestures, it significantly increases the awareness of tongue positioning and thus the effectiveness of corrective feedback with the help of tongue position images/videos increases.

4.9 Interactive CAPT system design

When creating a system for computer-assisted pronunciation teaching, understanding the language learner's requirements and motivation are important in order to achieve any lasting

success. There has been limited research on how to automatically teach pronunciation, i.e. what teaching methods or exercises are effective under various different circumstances. Derwing et al. [48] discuss the challenges in pronunciation teaching in general (i.e. independent of automation). They also called for more in-depth research around questions such as intelligibility, functional load and lasting impact of different pronunciation teaching approaches. Strik and al. [63] present a discussion of issues specific to design and pedagogy in the context of automated pronunciation teaching.

Language students have repeatedly expressed the desire to be told their pronunciation errors so that they know what to focus on. Accordingly, Neri et al. [49] showed that implementing corrective feedback even if in a limited form, does improve the pronunciation quality of students on an individual phoneme level and has a positive impact on user motivation.

The presentation layer of a CAPT system will greatly influence the acceptance rate of a system. Eskenazi et al. [5] and Yoon et al. [50] present some initial discussion of user interface design questions in conjunction with their own conclusions as to what to implement.

Sonu et al. [51] showed that both minimal pair based training and sentence level training is effective in order to improve a student's perception skills. Beyond that, there has been little work to incorporate pronunciation as part of dialog-based fluency training by engaging students in an interaction with a virtual tutor. A very early system that showed the feasibility of such an approach can be found in [52] and [53] where students had to traverse up to 7 states in order to complete an encounter. Likewise, Raux et al. [54] present an interesting approach to providing lessons within a dialog system by responding to ungrammatical sentences with a confirmation prompt that emphasizes the error. However, too high a non-native recognition error-rate prevented the effectiveness assessment of this approach.

Creating lesson material is a complex and time-consuming task. An attempt to help automating this process has been made in both [55] and [62]. Saz et al. [62] automatically identified confusable contexts that consist of an original sentence and an automatically generated sentence with a minimal pair difference. This can help students to focus on critical pronunciation errors that can cause a larger degree of misunderstanding than other pronunciation errors. There exists a large body of research on the lesson authoring for language learning, see for example Roseti et al. [56] who outline a multimedia authorship tool. Lastly, Johnson et al. [67] have built a lesson authoring system that incorporates both pedagogical considerations as well as Alelo's pronunciation teaching technology into new lessons. Alelo's products are one of the very few examples that have incorporate pronunciation learning in interactive multi-media dialog systems.

5 Existing Commercial Applications

Pronunciation error detection has two main commercial usages: (1) as part of pronunciation assessment and (2) as part of pronunciation teaching. Each application comes with a number of challenges, particularly on the pronunciation teaching side.

Product Name & Link	Company	Languages	Description
Versant and VersantPro	Pearson	English, Spanish, Arabic	Automated pronunciation assessment, measures speaking as well as listening
SpeechRater Engine (**http://www.ets.org/research/topics/as_nlp/speech**)	ETS	US English	Automated pronunciation assessment as part of standardized tests Pronunciation learning via AMEnglish.com includes training on stress, rhythm, intonation Part of TOELF since 2006
EnglishCentral	EnglishCentral	US English	English learning website, Assigns pronunciation score at sentence level, Tracks progress over time
CarnegieSpeech Assessment Climb Level 4 NativeAccent SpeakRussian SpeakFarsi	Carnegie Speech	Russian, Farsi	Pronunciation assessment as well as pronunciation teaching. Feedback at phone and sentence level Prosody?? Measurings pausing and duration
EduSpeak	SRI StarLab	Adults: American English, Latin American Spanish, French, German, Chinese (Mandarin), Arabic (Egyptian), UK English, Australian/NZ English, Japanese, Swedish, Tagalog (Filipino) Children: American English (Ages 4 to 15)	Acoustic modeling of childrens' speech
RosettaStone Totale.	RosettaStone (www.rosettastone.com)	Arabic, Chinese (Mandarin), Dari,Dutch, English (US,UK), Filipino, French, German, Greek, Hebrew, Hindi, Indonesian, Irish, Italian, Japanese,Korean, Latin, Pashto, Farsi, Polish, Protuguese, Russian,Spanish (LA and Spain), Swahili, Swedish, Turkish,Urdu, Vietnamese	Immersion approach, all teaching in target language
Spexx (www.speexx.com)	Digital Publishing	English, Spanish, French, Italian, German	Has 12 L1 language supports. Online programs, not software package. Also uses ASR for pronunciation training. Does word-level scoring with green.yellow, red highlighting).
TellMeMore v10.0 (www.tellmemore.com)	Auralog	Spanish, French,German, Italian, English, Dutch, Chinese, Japanese, Arabic	Front &sideview visualization of words, audio and F0 tracking.
EyeSpeak (www.eyeSpeakEnglish.com)	EyeSpeak	US and British English	Audio comparison, measures each phoneme, timing, loudness. Student can listen to each phoneme segment, visual cross-section of mouth for each sound, pitch tracking
Tactical Iraqi, Dari and Pashto	Aleho (alelo.com)	Iraqui, Dari and Pashto	Pronunciation teaching and immediate corrective feedback embedded in interactive, 3D video games.

Table 2: Summary of existing commercial CAPT systems

In the area of automated language skill assessment or pronunciation assessment, there has been quite some success in bridging the common gap between research and commercial development. The "phonepass" product suite from Pearson (formerly Ordinate) that is based on analyzing about 10 minutes of audio has been shown to measure pronunciation skills as reliably as human judges, see Bernstein et al.

[57],[58],[59]. Additionally, ETS (Educational Testing Services) and Pearson utilize complex algorithms to measure the pronunciation quality of students based on analyzing up to 10 minutes of speech. It has been shown that such algorithms can assess the pronunciation quality of a student as reliable as a trained human expert; see for example [58].

Several commercial language learning software packages have automated pronunciation error scoring tools incorporated. For example TellMeMore (previously Auralog), has exercises that allow the student to record their utterances, they display the wavefile recording and F0 contour, so that the student can compare those to the master recording and F0 contour. This exercise also gives the student a score for the entire word, but there is no phoneme level error feedback. Also, this exercise is L1 independent. A similar approach is being used in RosettaStone as well as EyeSpeak and Speexx. EyeSpeak has an interesting feature that displays the tongue position of the student in contrast to the teacher. In summary, recording a student's wavefile and allowing the student to compare their wavefile and F0 contour in terms of stress and duration with a teacher's example is a well-established practice in commercial systems. However, the challenges lie with the reliability of the judgment scale, especially when it comes to accented speech that is close to native speech.

6 Challenges in pronunciation error detection

The advantages of automated assessment are that CAPT could potentially be more reliable than human assessment, cheaper, and typically available any time and any place.
Based on the CAPT research overview in the previous section and data from an informal survey of leading researchers in this field, a list of core challenges has been:

1. Reliable phoneme-level error detection
2. Distortion error assessment
3. Text independence
4. L1 independence
5. Integrated assessment of both phonemic and prosodic pronunciation components
6. Corrective audiovisual feedback
7. Robust, interactive system design

The following paragraphs describe these seven challenges in more detail.

6.1 Reliable phoneme-level error detection

As observed in several papers, see for example [7], [60], the reliability of pronunciation error detection by human experts on the level of individual phonemes is fairly low. Likewise, automated error detection is not that reliable either, consequently the correlation between the two is even less. But in order to give productive feedback about the type of error that has been made, a CAPT system needs to identify the exact error within a word. On the other hand, false positives (telling a student there was an error when there was none) are potentially quite damaging for a language student.

6.2 Distortion error/accent detection

There has been little work on partially mispronunciations that is identifying error sources that contribute to perceived accent or on quantifying different degrees of accent.
Especially when the accented speech gets close to the target speech, the measurement uncertainty in existing algorithm is essentially larger than the actual difference between accented and target speech. For example, Mueller et al. [60] applied phoneme-level loglikelihood scoring to speech from near native language learners. It was found that the GOP score didn't correlate with human ratings. A very similar challenge has been addressed by Yan et al. [61], who used discriminative training to increase the quality of the acoustic models for similar phonemes such as 'sh-s' or 'n-l' and resulted in a small relative improvement.

6.3 L1-independence

As can be seen from section 4, there has been limited work that is L1-independent and yet has similar performance to methods based on knowledge of L2. On the other hand, the cost of non-native database collection and annotation is very high and does not scale. The challenge is to develop methods that derive a set of likely errors for a given student either from knowing his native language without requiring an annotated database for this L1/L2.

6.4 Text-independence

Conversational language training exercises requires text-independence. This can be achieved by improved adaptation and non-native acoustic modeling followed by forced-alignment pronunciation evaluation.

6.5 Integrated assessment

Most early work on pronunciation error detection has either focused on segmental errors or on suprasegmental features. However, especially for intermediate and advanced language learners, the majority of the errors are actually occurring with regard to the prosody. Prosody errors in particular tend to contribute to any perceived accent more so than individual phoneme mispronunciations.

6.6 Corrective audiovisual feedback

As can be seen from the limited availability in commercial language learning software, providing corrective feedback for pronunciation errors is a difficult task. There exist several systems that display cross-sections of the vocal tract in the attempt to show students how to move the tongue for each sound. However, the main challenge lies in making such illustrations effective so that the student knows how to implement the instruction into their own tongue movements.

6.7 Robust, interactive CAPT system design

As the review of the literature has shown there exist a large body of diverse research on almost any aspect of CAPT. The big challenge lies in combining all effective, appropriate methods in an integrated, interactive system that provides a comprehensive suite of exercises combined with trending.

7 Conclusion: Where we need to go

This paper summarized existing research on automated pronunciation error detection as well as some of the work on automated pronunciation error correction. The remaining challenges that need to be overcome in order to be able to develop truly useful pronunciation teaching applications have also been discussed.

Altogether, many components required for such applications already exist. However, one of the largest remaining challenges is to integrate these many components into one that ideally is L1-independent, or at least easily configured for a different L1, without requiring a manually annotated non-native database. It has been shown that many different features have been used to measure the various components of pronunciations. Thus future work may lie in combining more of such features in order to achieve a larger degree of accuracy and reliability in phoneme error detection as well as the detection of subtle degrees of fluent but accented speech.

Ideally such applications would utilize an intelligent virtual tutor that takes the role of a private tutor for the student. Such a tutor would have the means of providing corrective audio-visual feedback, including illustrating the differences between a student's pronunciation and that of a reference speaker, in a manner that helps students to figure out how to reduce their accent.

8 Acknowledgements

Many thanks to Jared Bernstein, Catia Cucchiarini, Farzad Ehsani, Maxine Eskenazi, Horacio Franco, Go Kawai, Yoon Kim, Antoine Raux, Helmer Strik, and Klaus Zechner for sharing their knowledge and assessment of the current state-of-the-art in pronunciation error detection.

9 References

[1] Catia Cucchiarini, Febe de Wet, Helmer Strik, and Lou Boves. "Assessment of dutch pronunciation by means of automatic speech recognition technology", ICSLP 1998, Sydney, 1998.

[2] Catia Cucchiarini, Helmer Strik, Lou Boves. "Automatic pronunciation grading for dutch", Proc. STiLL 1998, Stockholm, 1998.

[3] Catia Cucchiarini, Helmer Strik, and Lou Bouves. "Quantitative assessment of second language learners' fluency: An automatic approach", Jor. Acous. Soc., vol. 107, 1998.

[4] Maxine Eskenazi. "Using automatic speech processing for foreign language pronunciation tutoring:some issues and a prototype", Language Learning & Technology, 2:62–67, 1999.

[5] Maxine Eskenazi, Yan Ke, Jordi Albornoz, and Katharina Probst. "The fluency pronunciation trainer: Update and user issues", Proceedings INSTiL2000, 2000.

[6] Horacio Franco, Victor Abrash, Kristin Precoda, Harry Bratt, Ramana Rao, John Butzberger, Romain Rossier, and Federico Cesari. "The SRI EduSpeak system: Recognition and pronunciation scoring for language learning", INSTiL 2000, 2000.

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

[8] L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub, "Automatic scoring of pronunciation quality", Speech Communication, vol 30, p. 83-93, 2000.

[9] Horacio Franco. "Combination of machine scores for automatic grading of pronunciation quality", Speech Communication, vol 30, p. 121-130, 2000.

[10] Silke M. Witt. "Use of Speech recognition in computer-assisted language learning", unpublished thesis, Cambrige Uni. Eng. Dept, 1999.

[11] Maxine Eskenazi, "An overview of spoken language technology for education", Speech Communication vol 51, p. 832-844, 2009.

[12] Rodolfo Delmonte. "Exploring Speech Technologies for Language Learning", http://www.intechopen.com/books/speech-and-language-technologies, June 2011.

[13] John Levis. "Computer technology in teaching and researching pronunciation", Annual Review of Applied Linguistics, 27:184–202, 2008.

[14] Marianne Celce-Murcia, Donna Brinton, Janet Goodwin. "Teaching Pronunciation: A reference for teachers of English to speakers of other languages," New York: Cambridge University Press, Cambridge, 1996.

[15] Antoine Raux and Tatsuya Kawahara. "Automatic intelligibility assessmentand diagnosis of critical pronunciation errors for computer-assisted pronunciation learning", ICSLP 2002, Denver, USA, 2002.

[16] Christos Koniaris, Olov Engwall. "Phoneme Level Non-native Pronunciation analysis by an Auditory Model-based Native Assessment Scheme", Interspeech 2011, Florence, Italy.

[17] Go Kawai, Keikichi Hirose. "A CALL system using speech recognition to teach the pronunciation of Japanese tokushuhaku", Proc. STiLL 1998, Marholmen, Sweden, 1998.

[18] Brian Mak, Manhung Siu, Mimi Ng, Yik-Cheung Tam, Yu-Chung Chan, Kin-Wah Chan, Ka-Yee Leung, Simon Ho, Fong-Ho Chong, Jimmy Wong, and Jacqueline Lo. "PLASER: Pronunciation learning via automatic speech recognition.", Proc. HLT-NAACL 2003 Workshop on Building Educational Applications using Natural Language Processing, pages 23–29, 2003.

[19] Catia Cucchiarini, Henk van den Heuvel, Eric Sanders, Helmer Strik. "Error selecton for ASR-based English pronunciation training in My Pronunciation Coach", Interspeech 2011, Florence, Italy, 2011.

[20] Tobias Cincarek, R. Gruhn, C. Hacker, Elmar Nöth, and S. Nakamura. "Automatic pronunciation scoring of words and sentences independent from the non-native's first language", Computer Speech & Language, 23(1):65–88, January 2009.

[21] Akinori Ito, Yen-Ling Lim, Motoyuki Suzuki, and Shozo Makino. "Pronunciation error detection for computer-assisted language learning system based on error rule clustering using a decision tree", Acoustical Science and Technology, 28(2):131–133, 2007.

[22] Hui Ye, Steve Young. "Improving the Speech Recognition Performance of Beginners in Spoken Conversational Interaction for Language Learning", Interspeech 2005, Lisboa, Portugal, 2005.

[23] Oscar Saz, Eduardo Lleida, and William Rodríguez. "Acoustic-phonetic decoding for assessment of mispronunciations in speakers with cognitive disorders", AVFA09, 2009.

[24] Olaf Husby, Åsta Øvregaard, Preben Wik, Øyvind Bech, Egil Albertsen, Sissel Nefzaoui, Eli Skarpnes, and Jacques Koreman. "Dealing with L1 background and L2 dialects in norwegian CAPT", SLaTE 2011, Venice, Italy, August 2011.

[25] Ambra Neri. "Segmental errors in dutch as a second language: How to establish priorities in CAPT", Proceedings of the InSTIL/ICALL Symposium, Venice, Italy, 2004.

[26] Wai-Kit Lo, Shuang Zhang, Helen Meng. "Automatic Derivation of Phonological Rules for Mispronunciation Detection in a Computer-Assisted Pronunciation Training System", Interspeech 2010, Makuhari, Japan, 2010.

[27] Alissa M. Harrison, Win Yiu Lau, Helen Meng, Lan Wang, "Improving mispronunciation detection and diagnosis of learners' speech with contect-sensitive phonological rules based on language transfer," Interspeech 2008, Brisbane, Australia, 2008.

[28] Alissa M. Harrison, Wai-kit Lo, Xiao-jun Qian, Helen Meng. "Implementation of an extended recognition network for

mispronunciation detection and diagnosis in computer-assisted pronunciation training", SLaTE 2009, Birmingham, England, 2009.

[29] Xiajun Qian, Helen Meng, Frank Soong, "On mispronunciation Lexicon Generation using joing-sequence Multigrams in Computer Aided Pronunciation Training (CAPT)", Interspeech 2011, Florence, Italy, 2011.

[30] Theban Stanley, Kadri Hacioglu, and Bryan Pellom. "Statistical machine translation framework for modeling phonological errors in computer assisted pronunciation training system", SLaTE 2011, Venice, Italy, August 2011.

[31] Jost van Doremalen, Catia Cucchiarini, Helmer Strik. "Automatic Detection of Vowel Pronunciation Errors Using Multiple Information Sources", ASRU 2009, Merano, Italy, 2009.

[32] Khiet Truong, Ambra Neri, Catia Cuccchiarini, Helmer Strik. "Automatic pronunciation error detection: an acoustic-phonetic approach", INSTiL 2004, Venice, Italy, 2004.

[33] Ingunn Amdal, Magne Johnsen, Eivind Versvik. "Automatic evaluation of quantify contrast in non-native norwegian speech", SLaTE 2009, Birmingham, England, 2009.

[34] Helmer Strik, Khiet Truong, Febe de Wet, Catia Cucchiarini. "Comparing different approaches for automatic pronunciation error detection", Speech Communications vol 51, p. 845-852, 2009

[35] Su-Youn Yoon, Mark Hasegawa-Johnson, and Richard Sproat. "Landmark-based Automated Pronunciation Error Detection", Interspeech 2010, Makuhari, Japan, 2010.

[36] N. Moustroufas, Vassilis Digalakis. "Automatic pronunciation evaluation of foreign speakers using unknown text", Computer Speech and Language 21, p. 219-230, 2007.

[37] Lei Chen, Klaus Zechner, and Xiaoming Xi. "Improved pronunciation features for construct-driven assessment of non-native spontaneous speech". NAACL 2009, 2009.

[38] Jared Bernstein, Jian Cheng, Masanori Suzuki. "Fluency changes with general progress in L2 proficiency", Interspeech 2011, Florence, Italy, 2011.

[39] Gina-Anne Levow. "Investigating Pitch AccentRecognition in non-native speech", ACL 2009, Singapore, 2009.

[40] Anne Bonneau, Vincent Colotte. "Automatic feedback for L2 prosody learning", http://www.intechopen.com/books/speech-and-language-technologies , June 2011.

[41] Hansjörg Mixdorff, Daniel Külls, Hussein Hussein, Gong Shu, Hu Guoping, and Wei Si. "Towards a computer-aided pronunciation training system for german learners of mandarin", SLaTE 2009, Birmingham, England, 2009.

[42] Hussein Hussein, Hue San Do, Hansjörg Mixdorff, Hongwei Ding, Qianyong Gao, Guoping Hue, Si Wei, and Zhao Chao. "Mandarin tone perception and production by german learners", SLaTE 2011, Venice, Italy, 2011.

[43] Olov Engwall, "Analysis of and feedback on phonetic features in pronunciation training with a virtual teacher", Computer Assisted Language Learning, Vol 25, No. 1, p. 37-64, February 2012.

[44] Oliver Jokisch, Hongwei Ding, and Rüdiger Hoffmann. "Acoustic analysis of postvocalic /l/ in chinese learners of German in the context of an overall perception experiment", SLaTE 2011, Florence, Italy, 2011.

[45] Stephen Bodnar, Bart P. de Vries, Catia Cucchiarini, Helmer Strik, and Roeland van Hout. "Feedback in an ASR-based CALL system for L2 syntax: A feasibility study", SLaTE 2011, Venice, Italy, 2011.

[46] Grazyna Demenko, Agnieszka Wagner, Natalia Cylwik, and Oliver Jokisch. "An audiovisual feedback system for acquiring L2 pronunciation and L2 prosody", SLaTE 2009, Birmingham, England, 2009.

[47] Slim Ouni. "Tongue gestures awareness and pronunciation training," Interspeech 2011, Florence, Italy, 2011.

[48] Tracey Derwing and Murray Munro. "Second language accent and pronunciation teaching: A Research-Based approach", TESOL Quarterly, 39(3):379–398, September 2005.

[49] Ambra Neri, Catia Cucchiarini, and Helmer Strik. "ASR-based corrective feedback on pronunciation: does it really work?", Interspeech 2006, Pittsburgh, USA, 2006.

[50] Su-Youn Yoon, Lei Chen, Klaus Zechner. "Predicting Word Accuracy for the automatic speech recognition of non-native speech", Interspeech 2010, Tokyo, Japan, 2010.

[51] Mee Sonu, Keiichi Tajima, Hiroaki Kato, Yoshinori Sagisak, "Perceptual training of vowel length contrasts of Japanese by L2 listeners: Effects of an isolated word versus a word embeded in sentences," Interspeech 2011, Florence, Italy, 2011.

[52] Jared Bernstein, Amir Naini, Farzad Ehsani, "Subarashii: Encounters in Japanese Spoken Language Education", Calico Journal, 1999.

[53] Farzad Ehsani, Eva Knodt, "Speech Technology in Computer-Aided Language Learning: Strenghts and Limitations of a New CALL Paradigm", Language Learning & Technology, 2, 1: 45-60, 1998.

[54] Antoine Raux and Maxine Eskenazi. "Using Task-Oriented spoken dialogue systems for language learning: Potential, practical applications and challenges'. InSTIL 2004, 2004.

[55] Liu Liu, Jack Mostow, and Gregory Aist. "Automated generation of example contexts for helping children learn vocabulary", SLaTE 2009, Birmingham, England, 2009.

[56] Adroaldo Guimaraes Roseti, Almir dos Santos Albuquerque, Vasco Pinto da Silva Filho, Rogerio Cid Bastos. "Multimedia Authorship tool for the Teaching of Foreign Languages and distance Learning in a Multiagent Environment, in 'Multi-Agent Systems – Modeling, Control, Programming, Simulations and Applications, Dr. Faisal Alkhateeb (Ed.), ISBN: 978-953-307-174-9, InTech, 2011.

[57] Jared Bernstein, Alistarit Van Moere, Jian Cheng. "Validating automated speaking tests", Language Testing, Vol 27, p 355-377, 2010.

[58] Jared Bernstein, Jian Cheng. "Logic, Operation, and Validation of the PhonePass SET-10 Spoken English Test", Language Testing vol. 27, July 2010.

[59] Jared Bernstein, Masanori Suzuki, Jian Cheng, and U. Pado. "Evaluating diglossic aspects of an automated test of spoken modern standard arabic", SLaTE 2009, Birmingham, England, 2009.

[60] Pieter Mueller, Febe de Wet, Christa van der Walt, and Thomas Niesler. "Automatically assessing the oral proficiency of proficient L2 speakers", SLaTE 2009, Birmingham, England, 2009.

[61] Ke Yan, Shu Gong. "Pronunciation proficiency evaluation based on discriminatively refined acoustic models", I.J. Information Tech. and Comp. Science, Vol. 3, No 2. ,www.mecs-press.org, March 2011.

[62] Oscar Saz, Maxine Eskenazi. "Identifying Confusable Contexts for Automatic Generation of Activities in Second Language Pronunciation Training," SLaTE 2009, Birmingham, England, 2009.

[63] Helmer Strik, Frederik Cornillie, Jozef Colpaert, Joost van Doremalen, and Catia Cucchiarini. "Developing a CALL system for practicing oral proficiency: How to design for speech technology, pedagogy and learners" SLaTE 2009, Birmingham, England, 2009.

[64] Hongyan Li, Shen Huang, Shijin Wang, Bo Xu. "Context-dependent Duration Modelling with Backoff Strategy and Look-up Tables for Pronunciation Assessment and Mispronunciation Detection," Interspeech 2011, Florence, Italy, 2011.

[65] Florian Hönig, Anton Batliner, Karl Weilhammer, Elmar Noeth. "Islands of Failure: Employing word accent information for pronunciation quality asessment of English L2 learners", SLaTE 2009, Birmingham, England, 2009.

[66] Florian Hönig, Anton Batliner, Elmar Nöth. "Automatic Assessment of Non-Native Prosody – Annotation, Modelling and Evaluation", ISADEPT 2012, Stockholm, Sweden, June 2012.

[67] Lewis Johnson. "Error Detecton for Teaching Communicative Competence", ISADEPT 2012, Stockholm, Sweden, June 2012.