Available online at www.sciencedirect.com
ScienceDirect
Speech Communication 51 (2009) 832–844
SPEECH COMMUNICATION
www.elsevier.com/locate/specom

An overview of spoken language technology for education

Maxine Eskenazi *

Language Technologies Institute, Carnegie Mellon University, 4619 Newell Simon Hall, 5000 Forbes Ave., Pittsburgh, PA 15213, United States

Received 26 June 2008; received in revised form 26 February 2009; accepted 9 April 2009

Abstract

This paper reviews research in spoken language technology for education and more specifically for language learning. It traces the history of the domain and then groups main issues in the interaction with the student. It addresses the modalities of interaction and their implementation issues and algorithms. Then it discusses one user population – children – and an application for them. Finally it has a discussion of overall systems. It can be used as an introduction to the field and a source of reference materials.

Keywords: Language learning; Automatic speech processing; Speech recognition; Speech synthesis; Spoken dialogue systems; Speech technology for education

1. Introduction

This paper reviews the research in many areas of spoken language technology for education and especially for language learning. It looks at the history of this field and its progression to the state-of-the-art today. The order of the sections reflects aspects of the interaction of a tutoring system with the student-user. The field is highly multidisciplinary. It benefits from knowledge in computer science, statistics and signal processing as well as in second language acquisition, cognitive science and linguistics. The literature reveals that a variety of names for this field has been used, such as Computer-Assisted Language Learning (CALL) and Computer-Assisted Language Technologies (CALT). This paper will use the term that has been employed to describe work in Spoken Language Technology for Education, SLATE (thus it will not cover the area of non-spoken language processing). We will review findings by researchers using spoken language technology for education. Specifically, they develop education applications using automatic speech processing, sometimes using natural language processing and/or spoken dialogue processing where the processing techniques are created or modified for this application. Thus, other types of research that use spoken language technologies as off-the-shelf black boxes are not included here. This type of work can be found at venues such as CALICO (Computer-Assisted Language Instruction Consortium) and EUROCALL (European Association for Computer-Assisted Language Learning) and in journals such as Language Learning and Technology. For references to research specific to second language learning, the reader is referred to the publications of the American Association for Applied Linguistics (AAAL), the book by Ellis (1997) and the paper by Ellis and Bogart (2007).

For the use of natural language processing, such as for text applications, the reader can refer to such sources as Intelligent Tutoring Systems at AAAI (Association for the Advancement of Artificial Intelligence) and International Conferences on Artificial Intelligence in Education (AIED), as well as ACL (Association for Computational Linguistics) and NAACL (North American Association for Computational Linguistics), Human Language Technologies (HLT) conferences and related workshops such as SIGDIAL (Special Interest Group on Spoken Dialogue), the Natural Language Processing for Education

* Tel.: +1 4122683858.
E-mail address: max@cmu.edu

0167-6393/$ - see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.specom.2009.04.005

group workshops and ISCA SLaTE (International Speech Communication special interest group on Spoken Language for Education) for more information.

The references included in this paper were chosen in part for their accessibility. The reader is encouraged to consult other papers by the same authors. Since the field is constantly growing larger, all of the research in the area could not be cited. For other references, the reader is encouraged to consult the publications in which the papers cited here appear and to make use of the bibliography of each cited paper. Some areas have sparse literature to our knowledge and are not included here. We have also not delved into the area of aids for the handicapped since the literature seems more abundant in the methodology of how to correct errors rather than how to detect them. The reader should note that many of the techniques used in non-native pronunciation detection could be used for handicapped speech as well.

This paper therefore groups main issues in the interaction with the student. After some history of the field, it addresses the modalities of interaction with the student and algorithms and issues specific to the spoken language technologies: automatic speech recognition (ASR) used in the detection and assessment of pronunciation errors; perception training and the use of talking heads; detection and correction of prosody errors; and the use of synthesis. Next it deals with one of the student populations: children's speech and systems mostly designed for children, the reading tutors. Finally it discusses overall systems: spoken dialogue systems and whole tutoring systems including those using games and commercial systems.

Early SLATE applications resulted most often in proofs of concept rather than complete systems serving real users. From the late 1990s to present, that paradigm has shifted. A significant part of the SLATE effort at present involves real world systems and is often tested on real users/students. Essential to the success of those real systems is the multidisciplinary nature of the teams of researchers, including: spoken language technologists, teachers, second language acquisition experts, statisticians, signal processing experts and cognitive scientists.

1.1. Origins of SLATE research

SLATE research originated in a combination of necessity and the desire to apply automatic speech processing to new, more challenging populations of speakers. Destombes at IBM France (Destombes, 1993), after learning that his daughter was deaf, set out in the late 1970s to create a system that she could use to learn to speak intelligibly. His software displayed pitch and intensity versus time in a series of game interfaces. Destombes noted that there was earlier work using hardware, not computers, centered on a visible speech cathode ray tube and on the oscilloscope used by Mme. Borel-Maisonny. At the time of Destombes' work, simple pitch displays were used and individual phone detection was in its very early stages. Destombes also recognized the *organizational* role of prosody and its important role in making speech comprehensible. Some other early work using automatic display and detection for the deaf and hard of hearing was carried out by Martony (1968) and by Nickerson and Stevens (1972). The reader can also find references to early work in the domain in (Delcloque, 2000).

Later Flege (1988) demonstrated the value of using visual aids to train speakers to produce correct vowels in a target language. This finding reinforced the idea that if speech errors could be detected and displayed automatically, and if appropriate corrective information could also be displayed, non-native speakers could use this to learn how to correct their errors.

In the early 1980s, speaker-independent automatic speech recognition (ASR) was emerging. But while it worked well for some speakers in limited conditions, it performed poorly for others. Spectrograms revealed that speakers with higher pitched voices *appeared* to have less acoustic information in their speech. That observation and poor state-of-the-art ASR scores lead researchers to focus on adult male speech during the 1980s. It was not until the very end of the 1980s that female speech was also processed. Children's speech, with its higher pitch, was deemed to be more challenging than female speech and was not addressed at that time. Russell et al. (1996) tackled this issue by first recording a database of children's speech and then using it to train a speech recognition system. The recognition results showed promise, so he incorporated the child-trained ASR into a reading tutor that presented text on the computer screen for children to read aloud. The tutor followed their progress, matching expected utterances to what the student actually said and gave them positive feedback when they read correctly and help when they made mistakes.

Bernstein concentrated on another neglected population. He first worked on the recognition of speech from congenitally and adventitiously deaf adults and from adults with relatively severe Cerebral Palsy (and resultant dysarthria) (Bernstein, 1977, 1986; Bernstein et al., 1984). From this population, he moved to another ASR outlier population – non-natives. He first worked on pronunciation scoring for non-native adults (Japanese speakers of English) (Bernstein et al., 1989) and then on scoring of spoken language proficiency for non-natives more generally (Bernstein and Franco, 1996). The techniques he developed here aimed at measuring the ability of the non-native to listen and speak in real time.

Through the end of the 1990s, automatic speech recognition (ASR) was the main language technology used for language learning systems. Many other technologies, centered on natural language processing, cropped up in the late 1990s and began to be integrated in actual systems in the early 2000s.

2. Modalities of interaction and SLATE-specific issues

This section groups the modalities of interaction and related algorithms and issues in automatic processing:

automatic speech recognition (ASR) used in the detection and assessment of pronunciation errors; perception training and the use of talking heads; detection and correction of prosody errors; and the use of synthesis.

2.1. Pronunciation: detection and assessment

Ever since ASR was first applied to language learning, the limits of this technology have had to be palliated in some way. Researchers have addressed the non-deterministic nature of ASR output by first assessing its strengths and limitations. Then they rely on smart engineering to enhance the former and avoid encountering the latter. Finally, to succeed in getting real users to accept software that incorporates ASR, the expectations of the user must be shaped to match the actual capabilities of the systems.

System designers are concerned about the effect of incorrect speech processing, and thus system feedback, on the user. At present, users generally accept that a language learning system that uses automatic speech processing will make some errors. There are two types of errors. The system can determine that the user pronounced a correct sound or word when in reality the sound was incorrect. This is called a *false positive*. On the other hand, the system can determine that the user pronounced an incorrect sound or word when in reality it was correct. This is called a *false negative*. In reporting error detection rates, measuring both of these types of errors is important. They reflect the final impression given to the user. It is generally accepted that telling a user she is wrong when she is right has a stronger negative impact on learning than telling her she is right when she is wrong (Bachman, 1990). For this reason system designers will try to keep false negatives as low as possible, knowing that this will, in turn, have the effect of increasing false positives. A good general reference on automatic speech processing is the book by Huang et al. (2004).

There are two main approaches to detecting variations in the speech signal that are linked to non-native speech. One searches for individual errors in basic skills such as the pronunciation of an individual phone (we will refer to this as *individual error detection*) and the other determines the overall impression of fluent speech (we will refer to this as *pronunciation assessment*). In practice, as mentioned in (Strik's, this issue) comparison of several of these approaches, pronunciation assessment will arrive at some global score through the use of either overall measures of phonetic and prosodic precision or of some weighted average of local phoneme scores (some overall measures that are used are: temporal features, articulation rate and segment duration). When measured over longer periods of time than the phoneme, they can produce results that can be compared to human judgment (Bernstein and Franco, 1996). Individual error detection, on the other hand, requires the calculation of a score at a local (phoneme for phonetics, syllable or word for prosodics) level for each phone that is pronounced.

2.1.1. Individual error detection

The ability to precisely determine where a user is making errors is an advantage in pronunciation training. By knowing precisely which segment is incorrect and having a database of corrective materials corresponding to each segment type, when a system detects an error, it can look up the appropriate corrective feedback and show it to the user to help her understand what to do differently. Without this feedback, she probably would not be able to detect or correct the error herself, making random attempts at different articulator positions. For example, if we know that the user was supposed to pronounce a /v/ at the beginning of a word and the system has decided that the user pronounced something else, then the system can display text, illustrations, play the sound and/or produce some other corrective media that demonstrate how to articulate a /v/. Knowledge of the user's native language (L1) is useful here since speakers of different L1s have different problems for any given target language (L2). Thus a speaker of Hindi and a speaker of Spanish should not be given the same corrective instructions. Corrective feedback has been shown to increase learning (Precoda et al., 2000; Neri et al., 2006). In contrast, when assessing the global fluency of a speaker, many factors come into play. They are combined to create an overall score, reflecting the listener's impression of fluency (Bernstein and Franco, 1996). In this case, the overall assessment score is validated by comparison to human judgment of how fluent the speaker is in L2.

Some of the earliest publications in both individual error detection and pronunciation assessment can be found in 1996 and 1997 (Eskenazi, 1996; Neumeyer et al., 1996; Witt and Young, 1997). The approaches differed in what was compared to the incoming non-native utterance (non-native or native L2 speech or both) and in the method of speech recognition (either forced alignment or unconstrained speech recognition). Much of this research relied on techniques such as non-native speech adaptation, phonetic processing, confidence measures and other speech and signal processing strategies. The papers referred to by these authors and others in this area are excellent sources of knowledge about these techniques as is the above mentioned book by Huang and colleagues.

There are several approaches for the detection of individual errors. Many are discussed in detail in (Strik et al.'s, 2007) paper. Work by Kim and colleagues (1997) compared the use of HMM-based log-likelihood scores to HMM-based log-posterior probability scores and found the latter were best correlated with human judgment although precision on individual phones was low. They also found that increasing the overall number of phone scores per speaker improved their results. Langlais et al. (1998) used the same techniques and found similar results for non-native speakers of Swedish.

One approach (Svenster et al., 1998) used raw HMM-based recognition scores directly from a system trained on native speech. When native and non-native utterances were processed with this recognizer, the authors found that

the lowest scores seemed to correspond to non-native speech.

Another approach, from the thesis by Witt (1999), is Goodness of Pronunciation (GOP), which uses confidence scores drawn from recognition results. This approach has been used in other research such as Mak et al. (2003) and Neri et al. (2006).

Another approach is to use Linear Discriminant Analysis (LDA) as well as decision trees and phonological features (Tsubota et al., 2002). Tsubota found that these results could be improved by performing speaker adaptation (preprocessing) and segment-input pair-wise verification (postprocessing). He used non-native speech from one fairly homogeneous population – Japanese speakers of English.

Trong et al. (2005) used LDA trained on a small number of features and showed that they could detect differences in features that could in turn be communicated to the user. Weigelt et al. (1990) used the same analysis on features such as, for example, energy rate of rise, zero crossing rate and other measures in order to detect the difference between phonetic features. Phonetic features could be shown to students in the place of phones. But if phonetic features are shown individually, rather than being agglomerated into phones before feedback is given, then before learning the pronunciation of the phones, the user will need to be trained in the phonetics of the target language. There is some recent evidence (Liu et al., 2007; Yoshimura and MacWhinney, 2007) that teaching both phonetics and pronunciation enhances the uptake of pronunciation information when learning skills such as articulatory phonetics and grapheme–phoneme relations.

In order to assess these methods' precision in error detection, human judgment, that is, having some human experts (teachers, linguists, etc.) listen to, and annotate, the non-native speech, is used as the gold standard to which the automatic results are compared. This will be discussed further in the section on pronunciation assessment.

When using GOP, LDA and like approaches, the incoming speech can be compared to native speech, to both native and non-native speech, or to non-native speech. When dealing with non-native speech, there are several considerations. First, collecting non-native speech may be more difficult than collecting native speech. The native language of the speaker is another consideration. A collection of speech from speakers who are all of the same native language will yield a tighter statistical representation than, say a collection that includes native speakers of Japanese, French and Russian all speaking English as their L2. Here, for example, some of these speakers will tend to substitute S for TH and some will use F for TH. A TH model from speakers of several L1s would be less precise than a model of just the Japanese S for TH, for example. Furthermore, in our own unpublished work, we have observed that there is also a difference in the language learning level. While Japanese natives who are beginning learners of English may indeed almost all substitute S for TH, as they get more proficient, they will start to experiment with the production of other sounds to approximate the TH. Thus while many beginners' errors may be due to the influence of L1 and also to overgeneralization, regular explanations or models of the errors of the more advanced learner may not cover all possibilities. So the approaches which do not rely on predicting a closed set of possible sounds may succeed better at error detection for the more advanced learner (Bonaventura et al., 2000). When using non-native speaker data, it may be advantageous to divide the data by the L1s of the speakers and by their learning level (although this would require much computing power), or to use only speakers from one same L1 and level.

Some have tried to predict the substitutions, additions and deletions that a student could make mostly by using a set of rules to modify the pronouncing dictionary used for recognition to include all possible pronunciation variants that a student could produce for each lexical entry. The variants could be learned automatically from labeled data or entered by hand based on linguistic knowledge (Bonaventura et al., 2000). Work such as Raux and Kawahara (2002) used this approach with students with a Japanese L1 only. They eliminated the student level issue by having only beginning students as their subjects. Work by Ito et al. (2005) used error rule clustering with a decision tree to represent the variants.

Some other issues to account for include the distance between the L1 of the student and the target L2. For example, if a pair of languages has the same writing system (use the same characters), as French and English, then some rules could predict errors that are linked to incorrect character-to-sound mapping. However, if a pair of languages has different writing systems, the student may not link characters from her L1 to target characters.

Minematsu et al. (2007) used another approach based on the structural representation of the individual distortions of vowel spectra and of the relationships amongst the vowels in the acoustic space. They detected outliers which correspond to errors on vowels, independently of the speaker. This method, like the ones above, performed best when used to represent the speech of speakers who have both the same L1 and L2.

The assessment of these approaches must take into account increased learning of the individual segments as well as the overall learning effect. While the latter is the end goal and requires well-structured learning studies with real students, the measures of correct detection of each segment can be carried out on a database of previously-collected non-native speech. An interesting set of criteria that can be used to this end are discussed in (Cucchiarini et al., 2007). They list: errors common to speakers of the same L1, errors that are perceptually salient, those that would potentially hamper communication, frequent errors and persistent errors. These errors are linked to both system performance and the pedagogical relevance of the items.

As we see in all of the research mentioned in this section, it is both the type of detection mechanism and the pedagogical relevance of the item that is detected (does it really help students improve their speech) that are important.

2.1.2. Pronunciation assessment

Pronunciation assessment employs some processes that are similar to those used for individual error detection. It is most often used to reflect a listener's impression of the student's oral production capabilities on a more global level. Rather than aiming at the correction of specific errors, akin to the basic skill building approach in language teaching, pronunciation assessment tends to be closer to the immersive teaching approach, offering an overall impression, according to a combination of segmental and suprasegmental criteria. The two approaches are quite complementary, for example, the former for training and the latter to assess the success of the use of knowledge gained from using the former and applied to conversational speech.

Pronunciation assessment is based on a set of algorithms that can assess the quality of the non-native speech as a human expert would. The speech is judged on its natural flow (use of pauses, rhythm, use of pitch, etc.), overall correctness of articulation and other criteria that give the listener the impression of fluent communication. Thus emphasis here is on the perception of the non-native speech as well as on its production and it concerns prosody as well as phonetics. The gold standard is the judgment of a set of human listeners who are often experts in some aspect of language learning, such as classroom teachers or heads of national testing groups. Several of these experts are asked to judge a large amount of non-native data according to a predefined set of criteria that the automatic system will also use. It is important to obtain a high degree of inter-rater agreement. This agreement is commonly measured using Cohen's Kappa coefficient (Cohen, 2009). This measure can also then be used to compare the automatic assessment to the human gold standard.

As mentioned above, some of the earliest research on pronunciation assessment was published by Bernstein et al. (1990), Bernstein and Franco (1996), Townshend et al. (1998). They created a human-scored gold standard of data made up of large amounts of telephone-based responses to five different types of questions designed to reflect conversational speech. Using an HMM-based recognizer trained on a mix of native and non-native speech, they formed expected response networks. The system produced scores by both analysis of correct/incorrect responses and by function approximation using statistics output from the recognizer.

Cucchiarini and colleagues (1998, 2000, 2002) have developed a system to automatically assess non-native pronunciation of Dutch. They have used techniques derived from the ones mentioned in the above section that they also use for individual error detection. When creating their gold standard for comparison, they have taken care to assess the types of ratings and the types of labels they employ. The labels they use are: overall pronunciation, segment quality, fluency and speech rate. As far as the experts are concerned it is interesting to note that they have found more variance between individual raters' labels than between groups of raters from different backgrounds such as phoneticians and speech therapists.

Kawai and Hirose (1998) used a bilingual phone recognizer with native-trained acoustic models of the learner's L1 and L2 to identify insertions, deletions and substitutions of L2 phones. Recognition results were combined with phonetics, phonology and pedagogical knowledge to point to mispronounced phones and offer corrective feedback.

Zechner et al. (2007) have devised an automatic scoring method for use in rating the TOEFL^{R} iBT Practice Online product. They use a recognizer trained on a combination of native (Broadcast News) and non-native speech that produces word identity, timing and confidence scores. The system also generates features that are related to the speaker's perceived fluency. Scoring is again based on best fit to their human gold standard.

Rhee and Park (2004) used two model sets as well as noise-robust compensation, phonetic alignment and knowledge-based acoustic–phonetic parameter estimation. And Moustroufas and Digalakis (2007) rely on models of the native language (Greek) of the subjects to create a system that assesses utterances with no prior knowledge of linguistic content.

The definition of fluency varies slightly from author to author. Some of the indicators of fluency are the use (placement) and the frequency of pauses and the regularity of pacing. But these acoustic measures only gain relevance to the task at hand when they can be combined with other measures such as phonetic accuracy and word choice. In the former case, this avoids giving a high score to someone who speaks quickly but is impossible to understand. In the latter case, it separates those who are fluent when they "play it safe", using a small vocabulary, from those who can easily incorporate a variety of words on the fly.

2.2. Perception and talking heads

Historically, perception of L2 speech has been addressed by systems that did not use automatic speech processing. They simply presented prerecorded speech and the student was to write down what she heard. But the recent use of automatic methods in this area has obtained impressive results and demonstrated the utility of automatic speech processing for perception training. The automatic techniques used here involve resynthesis methods that enhance or somehow change the acoustic characteristics of phonetic segments (and sometimes suprasegmentals as well). The new acoustic version is played to the student, either in a series of gradual changes from the original speech to the new version, or in a set of comparisons of the new version and the original speech. The results are very successful for

perception and have sometimes also been shown to also have an effect on the students' production.

Akahane-Yamada et al. (1997, 2004), Bradlow et al. (1997), Pruitt et al. (1998), Akahane-Yamada and Adachi (1998) addressed the case where segments that are allophones in L1 are two perceptually distinct phones in L2. Using the case of Japanese listeners' comprehension of the R/L distinction in English, they believed that contrasting the natural sounds alone would have no effect on learning to distinguish them. Rather, they used the STRAIGHT software mentioned in the synthesis section to increasingly enhance the acoustic differences between the two sounds until they were perceived as two distinct entities. Results showed that this learned distinction was retained over long periods of time and, without supplemental phonetic or articulatory training, many students' productions of these sounds were also improved. This technique has been incorporated into a commercial language learning product by the ATR laboratories (www.atr.jp/html/product/product.html).

Instead of modifying the overall acoustics of a segment, Hazan and Simpson (1998) manipulated the acoustic cue that differentiates one segment from another. This draws attention to the specific difference between the segments and is effective in getting the students to attend to this difference since they would not notice it on their own. Cues relating to voicing, manner and place were modified and subjects perceived the difference between the segments. In later work (Hazan et al., 2005) added visual training to the audio training and obtained positive results for both perception and production of the sounds.

Many researchers have used visual information of many forms, from the waveform to the pitch contour, to guide students toward correct articulator placement and segment perception. Some of the most effective visual information has come from software that can automatically show, on a "talking head", the placement of the articulators and their movement for a given sound and from one sound to another for any given utterance in the target language.

One of the pioneers in this area is BALDI's creator, Massaro. His talking head shows continuous speech produced from several points of view (Massaro and Cohen, 1998; Massaro, 2003, 2006). The student can see a frontal view with the "skin" on, as a conversational agent, or with the "skin" removed to show the articulators. The head can be rotated to the side or to the back for a better view of the movement in certain articulations. It has been used in several language learning systems to teach reading, vocabulary and speech for several L2s (Massaro et al., 2006) and has also been used for tutoring children with hearing loss. It will automatically show the articulation of any L2 input graphemic or phonetic string.

Beskow et al. (2000) developed a series of talking heads whose facial characteristics could be modified to study the contribution to speech perception of the information coming from facial expressions such as raised eyebrows. They also studied the type of visual information used by students in pronunciation training, and the role of interactive agents in language learning.

Badin et al. (1998) also developed a talking head. This one was used to illustrate sounds that were presented in a manner that underlined each sound's relationship to every other sound in articulatory space.

The paper by Granstrom (2004) contains an excellent discussion of virtual tutors and their components (using talking heads) within the perspective of a system created at KTH in Stockholm.

2.3. Prosody detection and correction

While the correction of segmental errors helps a non-native speaker to be more understandable, someone with very good articulation may still be difficult to understand if her prosody is wrong. Prosody is the backbone of speech, providing the structure that links the individual sounds to one another and to the linguistic substrate. Thanks to this structure, a listener can predict what will come next. Good prosody can offset mediocre articulation. Yet real time pitch detection is not perfectly reliable. Furthermore, due to the elimination of information about individual variability in state-of-the-art recognizers, pitch information must be obtained from another type of analysis. This is not the case, however, for timing information which recognizers, particularly in forced alignment mode, furnish with high accuracy.

Pitch and duration detection compare a student's speech to that of a native speaker. Although it is fairly easy to create a composite representation of the variations in duration independently of speaking rate, it is much harder to create a composite of pitch variations and so, for the latter, students' utterances are often matched to the *one* closest (by some measure) native utterance.

Sundstrom (1998) used a speech recognizer to label incoming student speech and then to align it with a teacher's correct pronunciation. After taking pauses into account, the student's speech was modified in duration and F0 and resynthesized using PSOLA. This allows the student to practice improving her prosody while hearing a modified version of her own voice. It has been shown (Probst et al., 2002) that imitating a voice that is as close as possible to one's own is most effective for learning.

Delmonte (1998, 2000) took a two-level approach (syllable-word and utterance-phrase) to prosody learning in his SLIM system. He segmented the incoming speech and then aligned it with a native model and its transcription and gave it a phonetic description. Two models were created, one a top-down model for syllable-timed languages and the other, a bottom-up model for stress-timed languages. The system took L2 phonology into account and was assessed according to a set of several prosody level criteria that could each be taught individually.

Yamashita et al. (2005) used a multiple regression model to predict proficiency using F0, power and duration, comparing non-native to native utterances. This word-level

method was assessed for novice Japanese learners of English. Ishi and Hirose (2000) also used linear regression and applied it to segmental durations of a non-native utterance to find the duration appropriate to the student's speaking rate. This enabled them to reliably distinguish between single- and double-mora phonemes.

Rather than cover all possible sources of prosodic variation, Hincks (2004) narrowed her field of study to the detection of liveliness in non-native Swedish speakers' oral presentations in English. WaveSurfer was used for F0 extraction. Then she defined liveliness as the standard deviation from the mean F0 calculated over 10-s windows of speech. She normalized for pitch dynamics and found reliable values to distinguish lively and less lively presentations from advanced learners. Due to the fact that there are many factors at many different levels that affect the pitch contour, many of which remain under study, this approach lessens the amount of variability. Thus a reliable result was obtained by limiting the scope of research to one aspect of non-native variation, at one language learning level (advanced) where the student should start to have ingrained habits and thus less unaccounted-for variation.

Much remains to be done in the area of prosody tutoring. And, as we have seen, the need for this type of training is as strong as for phonetic training. Recent advances in the use of prosody information to improve speech recognition accuracy give hope that techniques developed there may be adapted for language learning purposes.

2.4. Speech synthesis

As researchers started to develop the first language learning applications using speech recognition, some thought about also using speech synthesis. Most agreed that any synthetic voice they used in their software would be taken to be an exemplar of native speech and therefore imitated by the student. Due to this, it is generally agreed that a voice in a SLATE application must be of very high quality. And so many system creators had decided to wait until better quality domain-independent synthesis was available. But a few have forged ahead to examine how state-of-the-art synthesizers can be used. Mercier et al. (2000) used Mbrola diphone concatenation for a bilingual spoken dictionary and student dictation exercises in their system built to teach Breton. Work by Seneff et al. (2004) described below uses ENVOICE (Yi and Glass, 1998) in spoken dialogues for a variety of learning tasks. A product called MathSpeak (MATHSPEAK, 2004) uses Cepstral LLC's synthesizer.

House et al. (1999) looked at the perception of synthesis by children, examining how children between the ages of 9 and 11 perceive and respond to prosodic variation. Both concatenative and formant synthesizers were used to vary both F0 and duration. They found that the prosodic differences were perceived and, when a "fun" voice is desired, the larger variations of F0 and duration were preferred.

Hincks (2002) used the WaveSurfer audiovisual synthesizer in a study of Swedish students learning technical English, specifically for pitch and duration differences in cognates. She was able to manipulate both pitch and duration in the tokens she presented to the students and observed long term learning of correct lexical stress.

Kawahara (2006) created the STRAIGHT speech synthesis system in a flexible format that allows it to be used as a research tool for L2 (we will discuss one of its uses in the perception section below). It allows easy manipulation of the speech signal to create emphasis, for example, on certain segments, thus drawing the student's attention to them. The manipulated segments can thus be presented in drills for perception and production tutoring.

Handley and Hamel (2005) have developed a benchmark for the evaluation of speech synthesis for language learning applications. They list the following requirements: quality of the output (comprehensibility as well as voice characteristics such as friendliness and expressiveness, the accuracy of pronunciation and its naturalness) and flexibility (the ability to adapt register, accent, etc., to the course material). They assessed the synthesis from three commercial products and divided their results by the type of skill being taught (reading pronunciation, conversation). Then they used three criteria, comprehensibility, acceptability and appropriateness, to grade each skill type. This principled approach demonstrates that there is no one-synthesis-fits-all solution for language learning and that using information from well-designed benchmarks will enable language learning system designers to choose the appropriate speech synthesis for their applications.

A good overview of the state of the art in speech synthesis can be found in (Black, 2007).

For basic work in areas mentioned in Section 2 of this paper, some tools that the reader may want to explore are the PRAAT software (http://www.fon.hum.uva.nl/praat/) and the Wavesurfer software (http://www.speech.kth.se/wavesurfer/) for spectrogram display.

3. A specific user population: children

This section concerns one specific set of users that presents a special challenge to speech technologies: children's speech. Since much of the work using children's speech in the literature concerns reading tutors, they are also addressed in this section.

Following Martin Russell's research on children's speech, a wealth of work has spring up both on recognizing children's speech for tutoring purposes and in using these results to create reading tutors. The interest in reading tutors has a dual origin. Increasing evidence of the importance of learning to read well at an early age has been a social motivator. From a scientific point of view, reading tutors paralleled the work on the automatic recognition of read speech (often centered on the use of the Wall Street Journal, BREF and other databases of read speech). In these systems, the linguistic content of what the child will

say is known ahead of time, thus making it easier to match the incoming speech to the model. However, recognition here has to deal with the young reader's characteristic hesitations, reprises and other variations which change the order of words and sounds from what would be expected.

Most research on children's speech starts with the collection of a significant amount of read speech from children who are not yet adolescents (whose voice characteristics have not yet changed). This data is either used to train a speech recognizer from scratch or to adapt the models of a recognizer trained on adult speech (most often female speech since it has higher pitch and comes from a shorter vocal tract, thus having characteristics closer to children's speech than male speech has).

Several authors have examined what makes children's speech different and how to automatically account for this difference. Russell et al. (1996), Qun and Russell (2001), D'Arcy and Russell (2005) and Gerosa and Giuliani (2004) have examined the overall acoustic properties of children's speech. Gerosa et al. (2006) have looked at segmental differences and Hacker et al. (2007) have looked at differences on the suprasegmental level. Hagen et al. (2007) proposed to use subword units to make recognition of children's speech for reading tutors more accurate.

Reading tutors first began to appear as prototype demonstrations of concept (Russell et al., 1996; Mostow et al., 1994). They showed both speech recognition capabilities and the ability to follow a young speaker reading despite reprises and other variations (noise, talking to friends, coughing, etc.). These systems were later put into real classrooms (Russell et al., 2000; Beck et al., 2004) where their effectiveness could start to be assessed.

Cosi et al. (2004) created a tutor for Italian students with disabilities and Mich et al. (2004) used their results mentioned above in a reading tutor for several languages. Cole et al. (1998) have addressed a variety of needs of young students such as those of the hearing impaired. Alwan and colleagues (2007) have assessed students' abilities to read aloud, using criteria that parallel the judgment that a teacher would make. There has also been a commercial venture by Soliloquy Learning.

4. Whole systems

This section concerns complete tutoring systems: spoken dialogue systems and whole tutoring systems including those using games and commercial systems.

4.1. Spoken dialogue systems

Attempting to furnish immersion-like learning opportunities, creators of ASR systems began to design limited short dialogues so that a system could conduct a fairly natural dialogue with the student.

Early dialogue systems for language learning spoke a prompt to the student and then displayed a selection of several possible responses on the screen for the student to speak. The answers were always very different from one another so that automatic recognition of any one of the answers was fairly reliable. At first only one of the possible answers was correct and each wrong response indicated a specific lexical or syntactic misunderstanding. This could be stored (or, in later systems, added to the student model) to be used in a correction module after the dialogue was completed. Subarashii (Ehsani et al., 1997, 2000; Bernstein et al., 1999) had a different approach. All of the responses were correct and each one lead the dialogue along a different path. Later versions of this system, based on Wizard of Oz (WOZ) studies (where a human imitates what the system's response would be) of what students actually say at a specific point in a dialogue, no longer offered multiple choice answers. They gave the student the impression that she could "freely" respond to a prompt and then matched the student's input to the closest amongst a set of utterances that the system expected to hear at that point in the dialogue. The prompt was designed to elicit only a limited number of responses.

Bianchi et al. (2004) used an ontology as the back end to their dialogue system. This provides rich and flexible material that can fuel a longer dialogue.

Researchers with experience in building robust dialogue systems have in recent years examined how to adapt their systems for learning. This involves modifying modules of the system, like the dialogue strategies employed at a specific turn. Seneff and her colleagues have been creating successful dialogue systems for many years. In the past few years they have applied this knowledge to language learning. Peabody et al. (2004) used their dialogue system without the speech components to assess the use of its natural language components to teach the lexical tones of Mandarin as displayed in Pinyin form. Seneff et al. (2004) have also made use of their multilingual dialogue experience in a system that is configured to support learning either Mandarin or English. Since the domain representation is independent of the target language, the latter can be changed while keeping the subject of the dialogues the same, such as hotel reservations. For conversational homework to learn Mandarin Chinese, using a spoken dialogue system, McGraw and Seneff (2007) have created ISLAND for narrow domains. An excellent summary of this group's efforts to create high quality dialogue systems (and games) for language learning can be found in (Seneff, 2007).

Others study the changes in dialogue strategies and their learning outcomes. Raux and Eskenazi (2004) use the CMU Olympus dialogue system architecture and the Let's Go bus information application for language learning by adding a new strategy. The system is given possible correct utterances for each point in the dialogue. As the speech from a student comes in, it is matched to the list of utterances. The system then takes the closest correct utterance and sends it to the synthesizer with markings of where it was different from the student's utterance. The synthesizer then enhances the different parts, thus acting like the human who will correct someone by emphasizing the cor-

rection, an action called *recasting* in the language learning literature. Forbes-Riley and Litman (2007), Forbes-Riley et al. (2007, 2008) have shown how detecting affect (uncertainty, emotion) and changing dialogue strategies appropriately has an effect on learning. Hollingsed and colleagues (2007) have also addressed affective state, looking at how to model sub-second level responsiveness by using machine learning and perceptual techniques. Liscombe and colleagues (2006) used acoustic and prosodic features (prosodic were the most useful) to detect question-bearing turns in tutorial dialogues. They have also examined other affective information for tutorial dialogue systems. VanLehn and his colleagues (2007) have studied the effects of dialogue strategies on learning for a long time and are now looking at the learning effects of a lower level of strategy that they call the turn-level tactic.

4.2. Whole tutoring systems

A whole tutoring system must present a curriculum of what skills the student is expected to learn, assess progress in some way, model the student so that it can progress through the curriculum in a principled manner, and furnish progress reports.

One of the earliest projects to create a complete system was (Hiller et al.'s, 1993) SPELL system that taught pronunciation (especially of vowels) to deaf children. The display of vowel space showed the ideal placement of the vowels and the placement of the vowel that the student uttered within that space. This, like Destombes' work, acknowledged the students' need for an appealing interface that could capture their interest.

An early system using HMM-based speaker-independent ASR was created by Rypa and Price (1999). Called the Voice Interactive Language Training System (VILTS), it was modeled on evidence from solid language learning research. The system covered many communicative skills necessary to the fluent speaker including: listening tasks with material representing a variety of speech types; reading practice that also served as pronunciation practice with topic selection and error detection; conversation activities that ranged from simple to complex with detection and feedback on fluency; and sustainment exercises in all of these areas intended to help the student maintain her skills.

The research at SRI in the 1990s gave birth to several systems, VILTS being one of the first. The work by Bernstein mentioned in the section on pronunciation assessment also came from this group. Another system from this group was created by Franco and colleagues (2000). Eduspeak was a toolkit for developers of language learning software that used state-of-the-art ASR and pronunciation scoring technology for multiple languages. Eduspeak offered system developers the means to create language learning exercises. They could change parts of the recognizer suite, such as the language model and the pronouncing dictionary to obtain tighter models of their application and thus better results.

LaRocca was another researcher to envisage a complete system early on. Like Rypa and Price, and a handful of others in the field, this system was based on his dual competency of having both technical knowledge and experience in language education. LaRocca et al. (2000) created systems that would detect spoken segmental errors and provide corrective feedback such as articulatory information in the form of a side view of a transparent head so that the student could see articulator placement.

Raux and Kawahara (2002) developed a system to tutor pronunciation for Japanese learners of English. It had an explicit model of the relationship between intelligibility and error rates (as opposed to pronunciation and error rates). This model was also used by their colleagues in a full interactive tutoring system endowed with a variety of exercises (Tsubota et al., 2004).

Mercier and colleagues (2000) created a system comprising a set of exercises to teach Breton to francophones. It taught both phonetic-to-orthographic mapping rules and pronunciation on the segmental and suprasegmental levels.

Cole and colleagues have produced tutoring systems for children for both the deaf student (Cole et al., 1999), the young reader (Wise et al., in press) and the non-native language learner (Cole et al., 1998). Their systems comprise rich interaction with a talking head and exercises prepared in collaboration with experts in each field. The group created an extensive database of children's speech to train their ASR.

4.2.1. Games

While speech technologies were maturing, another area has also come of age. Video games have developed very realistic and engaging interaction. Most computer users have seen and/or played these games at some time. This creates a challenge for language learning systems – can they engage the student and maintain their interest as well as these games can. This is especially critical for children who do not see the intrinsic value in learning certain subjects, and are thus not motivated, but willingly play computer games. The competitive nature of the game, be it competing with others or with oneself, is one aspect that has the power to hold the player's interest. There is little SLATE game software, as of this writing, but researchers are becoming increasingly engaged in creating interactive systems that have a game aspect to them. The field should see an explosion of work in this area in the near future.

One of the first to see the value of a game interface and invest the necessary effort in the graphic and curricular aspects of games is Johnson. He and his colleagues (2004), Mote et al. (2004) have created games designed to teach culture. With the first application aimed at American soldiers going to Iraq, he developed a series of scenarios in which the student plays a role, communicating with avatars. From this, the graphics were used for games to teach vocabulary and other linguistic knowledge. Recently this work has been commercialized (Johnson and Valente, 2008).

Chao and colleagues (2007) created a game for learning Chinese through translation. Students use repetitive translation as a means to internalize the vocabulary, grammar and pronunciation structures of Chinese. This activity was extended to include a dialogue game (Seneff et al., 2007) where the student carries on a dialogue using a variety of utterances to express the same linguistic content.

Wik and colleagues (2007) created the DEAL game using an avatar that plays roles to carry out a dialogue with the student, who has the impression of being able to freely express herself. Observational skills are called on and the student learns material at several levels: perception, vocabulary, grammar, etc.

4.2.2. Commercialized systems

The immense demand to learn other languages has provided the impetus for some of the prototype systems based on the research mentioned in these pages to be made into commercial systems. The creators are conscious of the need to make the capabilities of the technology correspond to the functions of the software, setting user expectations at the same level as system performance. In general each product has a specific goal and a well-defined set of skills to teach or to assess (or both). Systems teaching a skill will have a structured curriculum that is based on a set of skills to be learned and a progression through these skills, from basic to complex. There is a means to assess progress (detecting errors and keeping some sort of running account of them is one way to assess – giving a snapshot of student progress at any given time). And there is generally a report that can be used by the teacher and/or student to monitor progress. The systems are interactive, some are CD-based and some are delivered over the internet. There are many products that have been commercialized that imply the use of speech technologies. In fact few actually do use automatic speech processing. Following are a few that have roots in serious speech technology research.

To teach pronunciation, Carnegie Speech (www.carnegiespeech.com) developed NativeAccent, from the work on the Fluency project at Carnegie Mellon. Yamada's work on perception and production has resulted in a product commercialized by ATR (www.atr.jp/html/product/product.html). To assess the fluency of non-native speech, Ordinate, and then Harcourt and Pearson, has produced PhonePass, now known as Versant (www.ordinate.com/), based on the work by Bernstein that assesses fluency over the telephone.

Systems that tutor vocabulary and grammar as well as some pronunciation include Auralog (www.tellmemore.com/), Saybot (www.saybot.com/), and Rosetta Stone (www.rosettastone.com). For culture and language tutoring, Alelo (www.alelo.com) has commercialized courses such as Mission to Iraq. And the Soliloquoy system (Soliloquy Learning) teaches children to read aloud.

5. Conclusions

From modest beginnings, the use of automatic speech processing for education, and especially for language education, has blossomed in many directions. The techniques used have been honed and many have successfully been incorporated into tutoring systems. The market for tutoring software is large and should provide the continued momentum to push this field forward. Appealing systems that incorporate spoken dialogue and games are at the leading edge of the field. They will soon be central, providing not only tutoring, but also testbeds for the development of new algorithms and strategies.

While only a decade ago it was possible to cite most of the research in SLATE, the field is now large enough that we can only choose the most representative work. This paper attempts to cover research from many continents and countries on many topics. May other authors forgive me if their work is not covered in these pages. They will certainly be cited in the papers mentioned herein or they will have published in the venues that are cited. The reader should take this paper as a beginning point, suggestive of other sources of information.

Also missing from this paper is work on natural language processing, which makes up a large part of the software being developed for language learning. It is our hope that a similar review will appear on this subject.

Acknowledgement

The author would like to thank Dr. Alan W. Black for his comments and suggestions.

References

AAAL, <http://www.aaal.org>.

Akahane-Yamada, R., Adachi, T., 1998. Toward the optimization of computer-based second language production training. In: Proc. ESCA ETRW STiLL98, Marholmen, Sweden, pp. 111–114.

Akahane-Yamada, R., Adachi, T., Kawahara, H., 1997. Second language production training using spectrographic representations as feedback. J. Acoust. Soc. Jpn. E 18, 341–343.

Akahane-Yamada, R., Kato, H., Adachi, T., Watanabe, H., Komaki, R., Kubo, R., Takada, T., Ikuma, Y., 2004. ATR CALL: a speech perception/production training system utilizing speech technology. In: 18th Internat. Congress on Acoustics, Proc. ICA 2004, Vol. III, pp. 2319–2320.

Alwan, A., Bai, T., Black, M., Casey, L., Gerosa, M., Heritage, M. Iseli, M., Jones, B., Kazemzadeh, A., Lee, S. Narayanan, S., Price, P., Tepperman, J., Wang, S., 2007. A system for technology based assessment of language and literacy in young children: the role of multiple information sources. In: Proc. IEEE Multimedia Signal Processing Workshop, October 2007.

Bachman, L., 1990. Fundamental Considerations in Language Testing. In: Oxford Applied Linguistics. Oxford University Press, p. 216.

Badin, P., Bailly, G., Boe, L.J., 1998. Towards the use of a virtual talking head and of speech mapping tools for pronunciation training. In: Proc. ESCA ETRW STiLL98, Marholmen, Sweden, pp. 167–170.

Beck, J.E., Jia, P., Mostow, J., 2004. Automatically assessing oral reading fluency in a computer tutor that listens. Technol. Instruct. Cognit. Learning 2, 61–81.

Bernstein, J., 1977. Intelligibility and simulated deaf-like segmental and timing errors. In: Proc. IEEE ICASSP-77, pp. 244–247.

Bernstein, J., 1986. Applications of speech recognition technology in rehabilitation. In: Presented at an EIF Special Session on Speech Recognition, 1986 RESNA Meeting, Minneapolis, MN (Reprinted in Proc. Speech-to-Text Today and Tomorrow: A Conference at Gallaudet University, December, 1988).

Bernstein, J., Franco, H., 1996. In: Lass, N. (Ed.), Speech Recognition by Computer, Principles of Experimental Phonetics, Mosby, St. Louis.

Bernstein, J., Becker, R., Bell, D., Murveit, H., Poza, F., Stevens, G., 1984. Telephone communication between deaf and hearing persons. In: Proc. IEEE ICASSP-84, Paper 26.7.

Bernstein, J., Weintraub, M., Cohen, M., Murveit, H., 1989. Automatic evaluation of English spoken by Japanese students. In: Paper FF10, 118th Acoust. Soc. Amer. Meeting, November (Abstract in J. Acoust. Soc. Amer. 86 (Suppl. 1), S77).

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., Weintraub, M., 1990. Automatic evaluation and training in english pronunciation. In: Proc. ICSLP-90: 1990 Internat. Conf. on Spoken Language Processing, Kobe, Japan, pp. 1185–1188.

Bernstein, J., Najmi, A., Ehsani, F., 1999. Subarashii: encounters in spoken language education in Japanese. CALICO J. Special Issue Tutors Listen: Speech Recognition Language Learning 16 (3), 361–384.

Beskow, J., Granstrom, B., House, D., Lundeberg, M., 2000. Experiments with verbal and visual conversational signals for an automatic language tutor. In: Proc. ESCA ETRW INSTiL, Dundee, Scotland, pp. 138–142.

Bianchi, D., Mordonini, M., Poggi, A., 2004. Spoken dialog for e-learning supported by domain ontologies. In: Proc. ISCA ITRW INSTiL04, Venice, Italy, pp. 203–206.

Black, A., 2007. Speech synthesis for educational technology. In: Proc. ISCA ITRW SLaTE Workshop on Speech and Language Technology in Education, Farmington, PA.

Bonaventura, P., Herron, D., Menzel, W., 2000. Phonetic rules for diagnosis of pronunciation errors. In: Ilmenau, S. (Ed.), Konvens 2000, Tagungsband 5, Konferenz Verarbeitung natürlicher Sprache, pp. 225–230.

Bradlow, A., Pisoni, D., Akahane-Yamada, R., Tohkura, Y., 1997. Training Japanese listeners to identify English /r/ and /l/: IV. Some effects of perceptual learning on speech production. J. Acoust. Soc. Amer. 101, 2299–2310.

Chao, C., Seneff, S., Wang, C., 2007. An interactive interpretation game for learning Chinese. In: Proc. ISCA ITRW SLaTE, Farmington, PA.

Cohen, 2009. <http://en.wikipedia.org/wiki/Cohen's_kappa>.

Cole, R.A., Carmell, T., Connors, P., Macon, M., Wouters, J., de Villers, J., Tarachow, A., Massaro, D., Cohen, M., Beskow, J., Yang, J., Meier, U., Waibel, A., Stone, P., Fortier, G., Davis, A., Soldand, C., 1998. Intelligent animated agents for interactive language training. In: Proc. ESCA ETRW STiLL, Marholmen, Sweden, pp. 163–166.

Cole, R., Massaro, D.W., Rundle, B., Shobaki, K., Wouters, J., Cohen, M., Beskow, J., Stone, P., Connors, P., Tarachow, A., Solcher, D., 1999. New tools for interactive speech and language training: using animated conversational agents in the classrooms of profoundly deaf children. In: Proc. ESCA/SOCRATES ETRW on Method and Tool Innovations for Speech Science Education.

Cosi, P., Delmonte, R., Biscette, S., Cole, R.A., Pellom, B., van Vuren, S., 2004. Italian literacy tutor: tools and technologies for individuals wit cognitive disabilities. In: Proc. ESCA ETRW NLP Speech Technol. in Advanced Language Learning Systems Symposium, Venice, Italy, pp. 207–214.

Cucchiarini, C., Strik, H., Boves, L., 1998. Quantitative assessment of second language learners' fluency: an automatic approach. In: Proc. 5th Internat. Conf. on Spoken Language Processing (ICSLP'98), Sydney, Australia, pp. 2619–2622.

Cucchiarini, C., Strik, H., Boves, L., 2000. Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. J. Acoust. Soc. Amer. 107 (2), 989–999.

Cucchiarini, C., Strik, H., Boves, L., 2002. Quantitative assessment of second language learners' fluency: comparisons between read and spontaneous speech. J. Acoust. Soc. Amer. 111 (6), 2862–2873.

Cucchiarini, C., Neri, A., de Wet, F., Strik, H., 2007. ASR-based pronunciation training: scoring accuracy and pedagogical effectiveness of a system for Dutch L2 learners. In: Proc. Interspeech-2007, Antwerp, Belgium, pp. 2181–2184.

D'Arcy, S.M., Russell, M.J., 2005. A comparison of human and computer recognition accuracy for children's speech. In: Proc. ISCA Interspeech 2005, Lisbon.

Delcloque, P., 2000. History of CALL. <http://www.ict4lt.org/en/History_of_CALL.pdf>.

Delmonte, R., 1998. Prosodic modeling for automatic language tutors. In: Proc. ESCA ETRW STiLL, Marholmen, Sweden, pp. 57–60.

Delmonte, R., 2000. SLIM prosodic automatic tools for self-learning instruction. Speech Comm. 30 (2–3), 145–166.

Destombes, F., 1993. The development and application of the IBM speech viewer. In: Brekelmans, A., Elsendoorn, Ben A.G., Coninx, Frans (Eds.), Interactive Learning Technology for the Deaf (NATO ASI Series/Computer and Systems Sciences). Springer, pp. 187–198.

Ehsani, F., Bernstein, J., Najmi, A., Todic, O., 1997. Subarashii: Japanese interactive spoken language education. In: Proc. ESCA Eur. Conf. on Speech Communication and Technology. Eurospeech97, Rhodes, Greece, pp. 681–684.

Ehsani, F., Bernstein, J., Najmi, A., 2000. An interactive dialog system for learning Japanese. Speech Comm. 30 (2–3), 167–177.

Ellis, R., 1997. Second Language Acquisition. Oxford University Press.

Ellis, N.C., Bogart, P.S.H., 2007. Speech and language technology in education: the perspective from SLA research and practice. In: Proc. ISCA ITRW SLaTE, Farmington, PA.

Eskenazi, M., 1996. Detection of foreign speakers' pronunciation errors for second language training – preliminary results. In: Proc. Internat. Conf. on Spoken Language Processing (ICSLP), Philadelphia.

Flege, J.E., 1988. Using visual information to train foreign language vowel production. Language Learning 38 (3), 365–407.

Forbes-Riley, K., Litman, D., 2007. Investigating human tutor responses to student uncertainty for adaptive system development. In: Proc. Affective Computing and Intelligent Interaction (ACII), Lisbon, Portugal.

Forbes-Riley, K., Rotaru, M., Litman, D., Tetreault, J., 2007. Exploring affect-context dependencies for adaptive system development. In: Proc. Human Language Technol.: Ann. Conf. of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT), Rochester, NY.

Forbes-Riley, K., Litman, D., Rotaru, M., 2008. Responding to student uncertainty during computer tutoring: an experimental evaluation. In: Proc. 9th Internat. Conf. on Intelligent Tutoring Systems (ITS), Montreal, Canada.

Franco, H., Abrash, V., Precoda, K., Bratt, H., Rao, R., Butzberger, J., Rossier, R., Cesari, F., 2000. The SRI Eduspeak system: recognition and pronunciation scoring for language learning. In: Proc. ESCA ETRW INSTiL2000, Dundee, Scotland, pp. 123–128.

Gerosa, M., Giuliani, D., 2004. Preliminary investigations in automatic recognition of English sentences uttered by Italian children. In: Proc. ESCA ETRW NLP and Speech Technologies in Advanced Language Learning Systems Symposium, Venice, Italy, pp. 9–12.

Gerosa, M., Lee, S., Giuliani, D., Narayanan, S., 2006. Analyzing children's speech: an acoustic study of consonants and consonant–vowel transition. In: Proc. Internat. Conf. on Acoustics, Speech, and Signal Processing (ICASSP06), Toulouse, Vol. 1, pp. 393–396.

Granstrom, B., 2004. Towards a virtual language tutor. In: Proc. ISCA ITRW INSTiL04, Venice, Italy.

Hacker, C., Cincarek, T., Maier, A., Heßler, A., Noth, E., 2007. Boosting of prosodic and pronunciation features to detect mispronunciations of non-native children. In: Proc. IEEE ICASSP, Hawaii, Paper SLP-P3.8.

Hagen, A., Pellom, B., Cole, R., 2007. Highly accurate children's speech recognition for interactive reading tutors using subword units. Speech Comm. 49 (12), 861–873.

Handley, Z., Hamel, M.J., 2005. Establishing a methodology for benchmarking speech synthesis for computer-assisted language learning (CALL). Language Learning Technol. 9 (3), 99–119.

Hazan, V., Simpson, 1998. The effect of cue enhancement on consonant perception by non-native listeners: preliminary results. In: Proc. ESCA ETRW STiLL98, pp. 119–122.

Hazan, V., Sennema, A., Iba, M., Fuaulkner, A., 2005. Effect of audiovisual perceptual training on eh perception and production of consonants by Japanese learners of English. Speech Comm.

Hiller, S., Rooney, E., Laver, J., Jack, M., 1993. SPELL: an automated system for computer-aided pronunciation teaching. Speech Comm. 13, 463–473.

Hincks, R., 2002. Speech synthesis for teaching lexical stress. TMH-QPSR 44, 153–156.

Hincks, R., 2004. Processing the prosody of oral presentations. In: Proc. ISCA ITRW INSTiL, Venice, Italy, pp. 63– 68.

Hollingsed, T., Ward, N., 2007. A combined method for discovering short-term affect-based response rules for spoken tutorial dialog. In: Proc. ISCA ITRW Speech and Language Technology in Education (SLaTE), 2007.

House, D., Bell, L., Gustafson, K., Johansson, L., 1999. Child-directed speech synthesis: evaluation of prosodic variation for an educational computer program. Proc. ESCA Eurospeech 99, 1843–1846.

Huang, X., Acero, A., Hon, HW., 2004. Spoken Language Processing: A Guide to Theory, Algorithm and System Development. Prentice Hall.

Ishi, C.T., Hirose, K., 2000. Influence of speaking rate on segmental duration and its formulation for the use in CALL systems. In: Proc. ESCA ETRW INSTiL2000, Dundee, Scotland, pp. 106–108.

Ito, A., Lim, Y, Suzuki, M., Makino, S, 2005, Pronunciation error detection method based on error rule clustering using a decision tree. In: Proc. Interspeech 2005, Lisbon, Portugal, pp. 173–176.

Johnson, W., Valente, A., 2008. Tactical language and culture training systems: using artificial intelligence to teach foreign languages and cultures. In: Proc. IAAI 2008.

Johnson, W.L., Marsella, S., Vilhjálmsson, 2004. The DARWARS tactical language training system. In: Proc. I/ITSEC 2004.

Kawahara, H., 2006. STRAIGHT as a research tool for L2 study: how to manipulate segmental and supra-segmental features. In: Invited Talk at Fourth Joint Meeting of ASA and ASJ, December 2006. <http://www.wakayama-u.ac.jp/~kawahara/Resources/L2toolSTRAIGHT.pdf>.

Kawai, G., Hirose, K., 1998. A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training. In: Proc. Internat. Conf. on Spoken Language Processing – ICSLP-1998, Paper 0782.

Kim, Y., Franco, H., Neumeyer, L., 1997. Automatic pronunciation scoring of specific phone segments for language instruction. In: Proc. Eurospeech 97, Rhodes, Greece, pp. 645–648.

Langlais, P., Oster, A.M., Granstrom, B., 1998. Automatic detection of mispronunciation in non-native Swedish speech. In: Proc. ESCA ETRW STiLL, Marholmen, Sweden, pp. 41–44.

LaRocca, S., Bellinger, S., Potter, T., 2000. Voice-interactive German homework at the US Military Academy. In: Proc. ISCA ITRW INSTiL2000, Dundee, Scotland, pp. 26–30.

Liscombe, J., Venditti, J., Hirschberb, J., 2006. Detecting question-bearing turns in spoken tutorial dialogues. In: Proc. ISCA Interspeech 2006, Pittsburgh, PA.

Liu, Y., Massaro, D.M., Chen, T.H., Chan, H.L., Perfetti, C., 2007. Using visual speech for training chinese pronunciation: an in-vivo experiment. In: Proc. ISCA ITRW SLaTE, Farmington, PA.

Mak, B.S., Ng, M., Tam, Y., Chan, Y., Chan, K., Leung, K., 2003. PLASER: pronunciation learning via automatic speech recognition. In: Proc. HLT-NAACL, pp. 23–29.

Martony, J., 1968. On the correction of the voice pitch level for severely hard of hearing subjects. Amer. Ann. Deaf 113, 195–202.

Massaro, D.W., 2003. A computer-animated tutor for spoken and written language learning. Proc. 5th Internat. Conf. on Multimodal Interfaces (ICMI'03), Vancouver, British Columbia, Canada. ACM Press, New York, pp. 172–175.

Massaro, D.W., 2006. The psychology and technology of talking heads: applications in language learning. In: Bernsen, O., Dybkjaer, L., van Kuppevelt, J. (Eds.), Natural, Intelligent and Effective Interaction in Multimodal Dialogue Systems. Kluwer Academic Publishers, Dordrecht, The Netherlands, pp. 183–214.

Massaro, D.W., Cohen, M.M., 1998. Visible speech and its potential value for speech training for hearing-impaired perceivers. In: Proc. Speech Technology in Language Learning (STiLL'98), Marholmen, Sweden, pp. 169–172.

Massaro, D.W., Liu, Y., Chen, T.H., Perfetti, C.A., 2006. A multilingual embodied conversational agent for tutoring speech and language learning. In: Proc. 9th Internat. Conf. on Spoken Language Processing (Interspeech 2006 – ICSLP), Pittsburgh, PA, pp. 825–828.

MATHSPEAK, 2004. <http://www.gh-mathspeak.com/index.php>.

McGraw, I., Seneff, S., 2007. Immersive second language acquisition in narrow domains: a prototype ISLAND dialogue system. In: Proc. ISCA ITRW SLaTE, Farmington, PA.

Mercier, G., Guyomard, M., Siroux, J., Bramouille, A., Groumelon, H., Guillou, A., Lavanant, P., 2000. Courseware for Breton language spelling pronunciation and intonation learning. In: Proc. ISCA INSTiL2000, Dundee, Scotland, pp. 145–148.

Mich, O., Giuliani, D., Gerosa, M., 2004. Parling, a CALL system for children. In: Proc. ISCA ITRW NLP and Speech Technologies in Advanced Language Learning Systems Symposium, Venice, Italy, pp. 169–172.

Minematsu, N., Kamata, K., Asakawa, S., Makino, T., Hirose, K., 2007. Representation of pronunciation and its application for classifying Japanese learners of English. In: Proc. ISCA ITRW Speech and Language for Education (SLaTE) Workshop, Farmington, PA.

Mostow, J., Roth, S., Hauptmann, A.G., Kane, M., 1994. A prototype reading coach that listens. In: Proc. 12th Natl. Conf. on Artificial Intelligence (AAAI-94), Seattle, WA, pp. 785–792.

Mote, N., Johnson, W.L., Sethy, A., Silva, J., Narayanan, S., 2004. Tactical language detection and modeling of learner speech errors: the case of Arabic tactical language training for American English speakers. In: Proc. ISCA ITRW InSTIL04, pp. 47–50.

Moustroufas, N., Digalakis, V., 2007. Automatic pronunciation evaluation of foreign speakers using unknown text. Comput. Speech Language 21 (1), 219–230.

Neri, A., Cuccharini, C., Strik, H., 2006. ASR corrective feedback on pronunciation: does it really work? In: Proc. ISCA Interspeech, Pittsburgh, PA, pp. 1982–1985.

Neumeyer, L., Franco, H., Weintraub, M., Price, P., 1996. Automatic text-independent pronunciation scoring of foreign language student speech. In: Proc. Internat. Conf. on Spoken Language Processing, Philadelphia, PA.

Nickerson, R.S., Stevens, K.N., 1972. An experimental computer-based system of speech training aids for the deaf. In: Proc. Conf. on Speech Communication and Processing. Institute of Electrical and Electronics Engineers and Air Force Cambridge Research Laboratories.

Peabody, M., Seneff, S., Wang, C., 2004. Mandarin tone acquisition through typed interactions. In: Proc. ISCA ITRW INSTiL04, Venice, Italy, pp. 173–176.

Precoda, K., Halverson, C., Franco, H., 2000. Effect of speech recognition-based pronunciation feedback on second language pronunciation ability. In: Proc. ISCA ITRW INSTiL2000, Dundee, Scotland, pp.102–105.

Probst, K., Ke, Y., Eskenazi, M., 2002. Enhancing foreign language tutors – in search of the golden speaker. Speech Comm. 37 (3–4), 161–173.

Pruitt, J., Kawahara, H., Akahane-Yamada, R., 1998. Methods of enhancing speech stimuli for perceptual training: exaggerated articulation, context truncation and STRAIGHT resynthesis. In: Proc. ESCA ETRW STiLL98, Marholmen, Sweden, pp. 107–110.

Qun, L., Russell, M., 2001. Why is recognition of children's speech difficult? In: Proc. ESCA EUROSPEECH 2001, Aalborg, Denmark, pp. 2671–2674.

Raux, A., Eskenazi, M., 2004. Using task-oriented spoken dialogue systems for language learning: potential practical applications and challenges. In: Proc. ISCA ITRW INSTiL04, Venice, Italy, pp. 147–150.

Raux, A., Kawahara, T., 2002. Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning. In: Proc. 7th Internat. Conf. on Spoken Language Processing, Denver, Colorado, pp. 737–740.

Raux, A., Kawahara, T., 2002. Automatic intelligibility assessment and diagnosis of critical pronunciation errors for computer-assisted pronunciation learning. In: Proc. ICSLP 2002, Denver, pp. 737–740.

Rhee, S.C., Park, J.G., 2004. Development of the knowledge-based spoken English evaluation system and its application. In: Proc. ISCA INTERSPEECH 2004, pp. 1681–1684.

Russell, M., Brown, C., Skilling, A., Series, R., Wallace, J., Bonham, B., Barker, P., 1996. Applications of automatic speech recognition to speech and language development in young children. In: Proc. Internat. Conf. on Spoken Language Processing, ICSLP'96, Philadelphia, PA.

Russell, M.J., Series, R.W., Wallace, J.L., Brown, C., Skilling, A., 2000. The STAR system: an interactive pronunciation tutor for young children. Comput. Speech Language 14 (2), 161–175.

Rypa, M.E., Price, P., 1999. VILTS: a tale of two technologies, in tutors that listen: speech recognition for language learning, special issue. CALICO J. 16 (3), 385–404.

Seneff, S., 2007. Web-based dialogue and translation games for spoken language learning, keynote speech. In: Proc. ISCA ITRW SLaTE07, pp. 9–16.

Seneff, S., Wang. C., Zhang, J., 2004. Spoken conversational interaction for language learning. In: Proc. ISCA ITRW INSTiL04, Venice, Italy, pp. 151–154.

Seneff, S., Wang, C., Chao. C., 2007. Spoken dialogue systems for language learning. In: Proc. NAACL HLT07, Rochester, NY.

Soliloquy Learning, <www.soliloquylearning.com/>.

Strik, H., Truong, K., de Wet, F., Cucchiarini, C., 2007. Comparing classifiers for pronunciation error detection. In: Proc. Interspeech 2007, Antwerp, The Netherlands.

Strik, H., this issue. Comparing different approaches for automatic pronunciation error detection. Speech Comm.

Sundstrom, A., 1998. Automatic prosody modification as a means for foreign language pronunciation training. In: Proc. ESCA ETRW STiLL, Marholmen, Sweden, pp. 49–52.

Svenster, B., de Krom, G., Bloothooft, G., 1998. Evaluation and training of second language learners' pronunciation using phoneme-based HMMs. In: Proc. ESCA ETRW STiLL, Marholmen, Sweden, pp. 91–94.

Townshend, B., Bernstein, J., Todic, O., Warren, E., 1998. Estimation of spoken language proficiency. In: Proc. ESCA ETRW Speech Technology in Language Learning (STiLL), Stockholm, pp. 179–182.

Trong, K, Neri, A., DeWet, F., Cucchiarini, C, Strik, H., 2005. Automatic detection of frequent pronunciation errors make by L2 learners. In: Proc. ISCA Interspeech, Lisbon, pp. 1345–1348.

Tsubota, Y., Kawahara, T., Dantsuji, M., 2002. Recognition and verification of English by Japanese students for computer-assisted language learning system. In: Proc. ISCA Interspeech, pp. 1205–1208.

Tsubota, Y., Dantsuji, M., Kawahara, T., 2004. Practical use of autonomous english pronunciation learning system for Japanese students. In: Proc. ESCA ETRW INSTiL04, Venice, Italy, pp. 139–142.

VanLehn, K., Jordan, P., Litman, D., 2007. In: Developing Pedagogically Effective Tutorial Dialogue Tactics: Experiments and a Testbed, SLaTE Workshop on Speech and Language Technology in Education (ISCA Tutorial and Research Workshop), Farmington, PA.

Weigelt, L., Sadoff, S., Miller, J., 1990. The plosive/fricative distinction: the voiceless case. J. Acoust. Soc. Amer. 87, 2729–2737.

Wik, P., Hjalmarson, A., Brusk, J., 2007. DEAL a serious game for CALL practicing conversational skills in the trade domain. In: Proc. ISCA ITRW SLaTE, Farmington, PA.

Wise, B., Cole, R., van Vuuren, S., Schwartz, S., Snyder, L., Ngampatipatpong, N., Tuantranont, J., Pellom, B., in press. Learning to read with a virtual tutor: foundations to literacy. In: Kinzer, C., Verhoeven, L. (Eds.), Interactive Literacy Education: Facilitating Literacy Learning Environments Through Technology.

Witt, S.M., 1999. Use of Speech Recognition in Computer-Assisted Language Learning, Ph.D. Thesis. University of Cambridge, Dept. of Engineering, p. 445.

Witt, S., Young, S.J., 1997. Language learning based on non-native speech recognition. In: 5th Eur. Conf. on Speech Communication and Technology (Eurospeech 1997), September 1997, Rhodes, Greece, pp. 22–25.

Yamashita, Y., Kato, K., Nozawa, K., 2005. Automatic scoring for prosodic proficiency of English sentences spoken by Japanese based on utterance comparison. IECE Trans. Inform. Systems E88-D (3), 496–501.

Yi, J., Glass, J., 1998. Natural-sounding speech synthesis using variable-length units. In: Proc. ICSLP, Sydney, Australia, November 1998.

Yoshimura, Y., MacWhinney, B., 2007. The effect of oral repetition in L2 speech fluency: system for an experimental tool and a language tutor. In: Proc. ISCA ITRW SLaTE, Farmington, PA.

Zechner, K., Higgins, D., Xi, X., 2007. SpeechRater™: a construct-driven approach to scoring spontaneous non-native speech. In: Proc. ISCA ITRW Speech and Language for Education (SLaTE) Workshop, Farmington, PA.