PLASER: Pronunciation Learning via Automatic Speech Recognition

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**

Hong Kong University of Science and Technology
Department of Computer Science, and
Department of Electronic and Electrical Engineering
{bmak, eemsiu, ssmimi}@ust.hk

Abstract

PLASER is a multimedia tool with instant feedback designed to teach English pronunciation for high-school students of Hong Kong whose mother tongue is Cantonese Chinese. The objective is to teach correct pronunciation and not to assess a student's overall pronunciation quality. Major challenges related to speech recognition technology include: allowance for non-native accent, reliable and corrective feedbacks, and visualization of errors.

PLASER employs hidden Markov models to represent position-dependent English phonemes. They are discriminatively trained using the standard American English TIMIT corpus together with a set of TIMIT utterances collected from "good" local English speakers. There are two kinds of speaking exercises: minimal-pair exercises and word exercises. In the word exercises, PLASER computes a confidence-based score for each phoneme of the given word, and paints each vowel or consonant segment in the word using a novel 3-color scheme to indicate their pronunciation accuracy. PLASER was used by 900 students of grade 7 and 8 over a period of 2–3 months. About 80% of the students said that they preferred using PLASER over traditional English classes to learn pronunciation. A pronunciation test was also conducted before and after they used PLASER. The result from 210 students shows that the students' pronunciation skill was improved. (The statistics is significant at the 99% confidence level.)

1 Introduction

The phenomenal advances in automatic speech recognition (ASR) technologies in the last decade led to the recent employment of the technologies in computer-aided language learning (CALL) [1]. One example is the LISTEN project (Mostow et al., 1994). However, one has to bear in mind that the goal of ASR in most other common classification applications (such as automated call centers, dictation, etc.) is orthogonal to that in CALL: while the former requires ASR in general to be forgiving to allophonic variations due to speaker idiosyncrasies or accent, pronunciation learning demands strict distinction among different sounds though the extent of strictness could be very subjective with a human teacher. As a result, technologies developed for mainstream ASR applications may not work satisfactorily for pronunciation learning.

In the area of pronunciation learning, ASR has been used in CALL for two different purposes: teaching correct pronunciation of a foreign language to students (Kawai and Hirose, 2000), and assessing the pronunciation quality of a speaker speaking a foreign language (Witt and Young, 2000; Neumeyer et al., 2000; Franco et al., 2000). The former asks for accurate and precise phoneme recognition while the latter may tolerate more recognition noises. The judgment for the former task is comparatively more objective than that for the latter which, on the other hand, is usually required to correlate well with human judges. In this paper, we describe a multimedia tool we built for high-school students in Hong Kong to self-learn American English pronunciation. Their mother tongue is Cantonese Chinese. The objective is to teach correct pronunciation of basic English phonemes (possibly with local accent), and not to assess a student's overall pronunciation quality. Although there

[*] Mr. Tam is now a graduate student at the Department of Computer Science at Carnegie Mellon University.

[†] Mr. Chan is now working at SpeechWorks Inc.

[1] CALL applies many different technologies to help language learning, but this paper concerns only the one area of pronunciation learning in CALL.

exist commercial products for the purpose, they have two major problems: First, they are not built for Cantonese-speaking Chinese; and, second, the feedback from these products does not pinpoint precisely which phonemes are poorly pronounced and which phonemes are well pronounced. As a matter of fact, most of these systems only provide an overall score for a word or utterance. As the feedback is not indicative, students would not know how to improve or correct their mistakes. One reason is the relatively poor performance of phoneme recognition — the best phoneme recognition accuracy is about 75% for the TIMIT corpus.

We took a pragmatic view and designed a multimedia learning tool called PLASER — Pronunciation Learning via Automatic SpEech Recognition — according to our following beliefs and guidelines:

1. It is an illusive goal for average students to learn to speak a second language without local accent. Therefore, PLASER should be tolerant to minor Cantonese accents, lest the students become too frustrated from continually getting low scores. For example, there is no “r” sound in Cantonese and consequently Cantonese usually speaks the “r” phoneme with weak retroflexion.

2. Performance of phoneme recognition over a long continuous utterance is still far from being satisfactory for pedagogical purpose.

3. PLASER’s performance must be reliable even at the expense of lower accuracy.

4. To be useful for correcting mistakes, PLASER must provide meaningful and indicative feedbacks to pinpoint which parts of an utterance are wrongly pronounced and to what extent.

5. The knowledge of IPA symbols is not a pre-requisite to learning pronunciation.

This paper is organized as follows: in the next Section, we first present the overall system design of PLASER. This is followed by a discussion of our acoustic models in Section 3. Section 4 gives a detailed description of our confidence-based approach in pronunciation scoring, and the related feedback visualization is given in Section 5. Both quantitative and qualitative evaluation results are given in Section 6. Finally, we summarize the lessons we learned in building PLASER and point out some future works in Section 7.

Table 1: Phonemes that are taught in PLASER (written in TIMIT-bet)

Lesson#	Phoneme	Pair	Lesson#	Phoneme	Pair
1	iy	ih	11	k	g
2	eh	ey	12	s	z
3	ae	ah	13	sh	zh
4	aa	ao	14	ch	jh
5	ax	er	15	f	v
6	ow	uh	16	th	dh
7	uw	ay	17	m	n
8	oy	aw	18	ng	h
9	p	b	19	l	r
10	t	d	20	w	y

2 PLASER: System Design

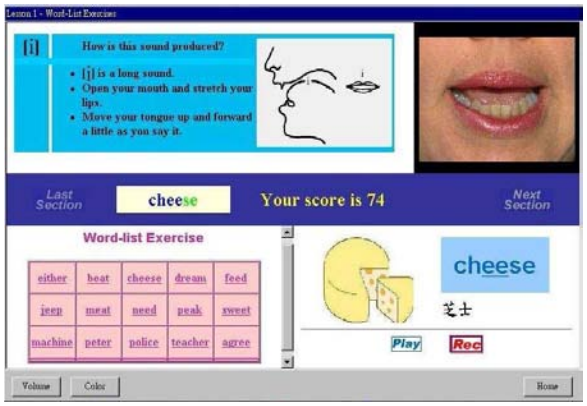
PLASER runs under Microsoft Windows (98, NT, 2000) with an easy-to-use web-like interface requiring only standard utilities such as the Internet Explorer and Media Player. PLASER consists of 20 lessons, and each lesson teaches two American English phonemes as shown in Table 1. The two phonemes in a lesson are usually the most confusable pair among the 40 phonemes. PLASER contains a lot of word examples and for each word there are its English spelling, its Chinese translation, a picture, and a pronunciation video-clip (PVC) which a native American English speaker helped record. A user may read and listen to the materials of each word as many times as he likes at his own pace. Besides descriptive materials, PLASER uses four types of exercises to teach pronunciation:

Read-Along Exercise: Basic pronunciation drills with no assessment.

Minimal-Pair Listening Exercise: This is used to train users’ ear. Words from one minimal pairs are randomly embedded in a sentence that makes perfect sense with either word in the pair. A user listens to recordings of such sentences and chooses between the two words.

Minimal-Pair Speaking Exercise: Similar to the Minimal-Pair Listening Exercise except that now only minimal pairs are given and a user is asked to say them. A student may pick any one of the two words to say but not to mix up with its counterpart in the pair. It is a two-class classification problem.

Word-List Speaking Exercise: A student may pick any word from a list to say, and PLASER has to decide how well each phoneme in the word is pronounced.



How is this sound produced?

- [i] is a long sound.
- Open your mouth and stretch your lips.
- Move your tongue up and forward a little as you say it.

Your score is 74

either	beat	cheese	dream	feed
jeep	meat	need	peak	sweet
machine	peter	police	teacher	agree

Figure 1: A snapshot of PLASER running its word exercise

Fig. 1 shows a snapshot of PLASER running the Word-List Speaking Exercise in the lesson teaching the two phonemes: “ih” and “iy”. The user has selected the word “cheese” to practise. The top left panel tells how to produce the phoneme “iy” with the help of an animated GIF that shows a cross-sectional view of the vocal tract during the phoneme’s production. At the bottom right panel are the word’s spelling, its Chinese translation, its picture, plus a recording button and a playback button. The word’s PVC is shown at the top right panel. The middle panel in the screen is reserved for feedbacks. The feedback for Word-List Speaking Exercise consists of an overall score for the practising word (“cheese” here) as well as a confidence score for each individual phoneme in the word using a novel 3-color scheme. Confidence scores are derived from a log-likelihood ratio between the desired target and some reference. Garbage rejection is also implemented in a similar manner. Refer Section 4 and 5 for more details.

As a self-learning as well as a teaching aid, the length of each lesson is designed to take about 25–30 minutes to complete. Students’ performance is recorded for later reviews by students themselves if PLASER is used as a learning tool, or by teachers if PLASER is used as a teaching aid.

3 Acoustic Modelling

For the development of PLASER’s acoustic models, additional speech data were collected from local high-school students:

HKTIMIT: A set of TIMIT utterances collected from a group of 61 local (Cantonese) high-school students who spoke “good” English to the local standard. There are 29 females and 32 males, and each recorded 250 TIMIT sentences. The data were divided into a training set of 9,163 utterances from 17 females and 20 males, and a test set of 6,015 utterances from 12 females and 13 males.

MP-DATA: A superset of words used in PLASER’s minimal-pair exercises recorded by eight high-school students, 4 males and 4 females, each speaking ∼300 words for a total of 2,431 words.

WL-DATA: A superset of words used in PLASER’s word exercises by the same eight students who recorded the MP-DATA for a total of 2,265 words.

All data were recorded with the same conditions as those of TIMIT. In addition, all utterances of MP-DATA and WL-DATA were phonetically transcribed.

The standard American English TIMIT corpus together with the HKTIMIT corpus were used to develop Cantonese-accented English phoneme HMMs. The common 13 mel-frequency cepstral coefficients and their first and second order derivatives were used for acoustic representation. All phoneme HMMs have three real states, and there are an additional 3-state silence model and a 1-state short-pause HMM. Three kinds of modelling techniques were investigated:

Context-Independent Modelling: Context-independent HMMs (CIHMM) were trained for the 40 phonemes taught in PLASER. Including the silence and short-pause models, there are totally 42 HMMs.

Position-Dependent HMM (PDHMM): Due to concerns of limited computing resources in local public schools, a restricted form of context-dependent modelling was chosen. Since PLASER will only perform phoneme recognition on isolated words, we postulate that it may be important to capture the word-boundary effect of a phoneme. Thus, three variants of each phoneme are modelled depending on whether it appears at the beginning, in the middle, or at the end of a word.

(MCE) Discriminative Training: With the goal of minimizing classification errors in a development data-set which is WL-DATA in our case, word-based MCE/GPD algorithm (Juang and Katagiri, 1992; Chou, 2000) was applied to improve the EM-trained acoustic models.

We started with a baseline system using 40 monophones with 24 mixtures per state. It gives a phoneme recognition accuracy of 39.9% on the HKTIMIT test set. The low accuracy perhaps indicates an unexpected lower

Modelling Technique	Classification Acc. of MP-DATA
CIHMM, 24 mixtures	81.50
CIHMM, 24 mixtures + MCE	84.48
PDHMM, 20 mixtures	82.83
PDHMM, 20 mixtures + MCE	85.29

Table 2: Investigation of various modelling techniques on minimal-pair classification

English proficiency of local students as well as a large deviation of local English from native American English. We then investigated PDHMM and MCE training, and gauged our progress by the classification accuracy of minimal pairs in the MP-DATA set. The results are tabulated in Table 2.

By using PDHMMs, the inventory of models is only increased by three times, requiring little additional computational resources. Yet they result in a relative error reduction of 7.2%. MCE discriminative training gives an additional relative improvement of about 14–16%.

4 Confidence-based Phoneme Assessment

The assessment of pronunciation accuracy is cast as a phoneme verification problem. The posterior probability of a phoneme is used as the Goodness of Pronunciation measure (GOP), which has been shown in many works (Witt and Young, 2000; Franco et al., 2000) that it is a good measure. PLASER computes both a GOP score and a normalized GOP score for two types of feedback as will be discussed in Section 5.

When a student runs a PLASER word exercise, s/he will randomly pick a word from a list and watches its pronunciation video-clip (PVC). When s/he feels comfortable to try s/he records her/his voice speaking the word. PLASER then computes a confidence-based GOP for each phoneme in the word as follows.

STEP 1: PLASER consults its dictionary for the standard phonemic transcription of the word which should be the same as that of its PVC.

STEP 2: Based on the transcription, forced alignment is performed on the student's speech.

STEP 3: For each acoustic segment X_u of phoneme y_u (where u denotes the phoneme index), PLASER computes its GOP(y_u), s_u, as its posterior probability by the following log-likelihood ratio normalized by its duration T_u:

$$\begin{aligned} s_u &= \log Prob(y_u|X_u) \\ &\approx \frac{1}{T_u} \cdot \log \left[\frac{p(X_u|y_u)p(y_u)}{\sum_{k=1}^{N} p(X_u|y_k)p(y_k)} \right] \quad (1) \\ &\approx \frac{1}{T_u} \cdot \log \left[\frac{p(X_u|y_u)}{p(X_u|y_{j_{max}})} \right] \quad (2) \end{aligned}$$

where N is the number of phonemes, and j_{max} is the phoneme model that gives the highest likelihood of the given segment. This GOP is used with some thresholds to decide if the phoneme is pronounced correctly.

In practice, the denominator in Equation 2 is replaced by the Viterbi likelihood of the segment given by a phone loop. Notice that the Viterbi path of a segment may contain more than one phoneme model.

STEP 4: Besides the raw GOP score, GOP(y_u) = s_u computed in STEP 3, a *normalized GOP score* is also computed by normalizing the GOP score to the range [0.0 .. 1.0] using a sigmoid function. That is, the normalized GOP for the phoneme y_u is given by

$$\text{sigmoid}(s_u) = \frac{1}{1 + \exp(-\alpha s_u + \beta)} \quad (3)$$

where the parameters α and β are empirically found.

The current PLASER implementation has some modifications due to practical reasons: The phone loop for computing the denominator of Equation 2 uses only the middle-position PDHMM of each phoneme plus the silence and short pause models for faster computation. For greater computation savings, the phone loop may also be replaced by a single Gaussian Mixture Model (GMM) trained by all phoneme segments in the training data. In our experience, a GMM with 32 mixtures suffices with a slight degradation in performance.

5 Visualization of Recognition Results

Two kinds of feedback of different resolutions are given for the word exercise:

- an overall phoneme score of the whole word; and,
- a phoneme-by-phoneme assessment by a 3-color scheme.

5.1 Overall Phoneme Score of a Word

The use of posterior probability as the GOP score for assessing the accuracy of a phoneme segment allows us to readily define an overall phoneme score (PS) for a word as a weighted sum of the normalized GOPs of its composing phonemes:

$$PS(\text{word}) = \sum_{k=1}^{N} w_k \cdot \text{normalized-GOP}(\text{phoneme}_k) \quad (4)$$

where w_k is the weighting of the k-th phoneme among the N phonemes composing the word. In the current PLASER, all phonemes in a word are equally weighted.

5.2 A 3-Color Feedback Scheme for Phoneme Confidence

The usefulness of an overall confidence for a word may be limited as it does not pinpoint the pronunciation accuracy of each phoneme in the word, and thus, the user still does not know how to correct his mistakes when the score is not good. Any attempt to report phoneme confidence score has to face the following two problems:

- unless users can read phonemic transcriptions, it is not clear how to report the confidence scores at phoneme level; and,
- unless the phoneme confidence scores are highly reliable, reporting its precise value may be too risky.

Our solution is a visual feedback that gives a color to the letters in the word spelling to indicate the pronunciation accuracy of their associated phonemes. To do that,

STEP 1: We first designed a rule-based algorithm to map each phoneme in the transcription of a word to its spelling letters. For example, for the word "beat" with the phonemic transcription "`/b/ /iy/ /t/`", the three phonemes are mapped to the letters "b", "ea" and "t" respectively. On the other hand, for the word "eve" with the phonemic transcription "`/iy/ /v/`", the two phonemes are mapped to the letters "e" and "v" respectively while the last letter "e" is not mapped to any phoneme.

STEP 2: A novel 3-color scheme was devised to reduce the preciseness of phoneme confidence scores. Two thresholds were found for each phoneme to label its confidence as good, fair, or bad. If the confidence score of a phoneme is good/fair/bad, its corresponding spelling letter(s) is/are painted in blue/green/red respectively. Two examples are shown in Fig. 2. The use of colors is also more appealing to users.

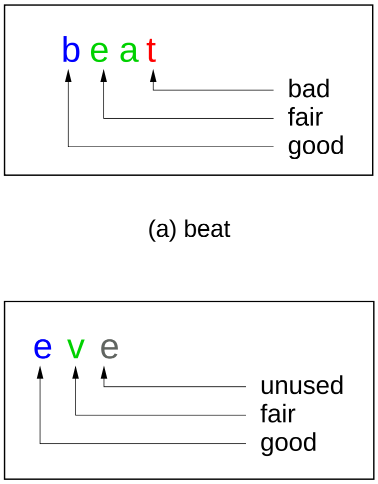
To find the two thresholds in the 3-color scheme, we treated the problem as a bi-threshold verification problem. The detailed algorithm is beyond the scope of this paper and will only be briefly described here. For details, please refer to (Ho and Mak, 2003).

Firstly, one has to decide how forgiving one wants to be and specifies the following two figures:

- the false acceptance rate (FA) for an incorrectly pronounced phoneme; and,
- the false rejection rate (FR) for a correctly pronounced phoneme.

If one sets FA very low, it will be hard to get "blue" scores; on the other hand, if one sets FR very low, it may be too forgiving and "red" scores will rarely show up. Due to its bi-threshold nature, it turns out that in such circumstances, simple method to determine the two thresholds will results in dominating "green" scores with little "blue" or "red" scores. The more complicated algorithm in (Ho and Mak, 2003) tries to avoid that.

Furthermore, due to scarcity of training data in the development data set, the phonemes were grouped into 9 phoneme classes in PLASER, and class-dependent thresholds were determined from the development data set. The 9 phoneme classes are: affricates, diphthongs, fricatives, nasals, semi-vowels, stops, back vowels, mid vowels, and front vowels.



(a) beat

(b) eve

Figure 2: A three-color scheme for showing phoneme confidence (The figure has to be read with color printouts, or electronically on a color display. The letters marked with "bad", "fair", "good", and "unused" are painted in red, green, blue, and gray respectively.)

6 Evaluation

A beta version of PLASER was tested by 900 students of Grade Seven and Eight over a period of about 3 months in twelve high schools. Both quantitative and qualitative evaluations were conducted to gauge the effectiveness of using PLASER to learn English pronunciation.

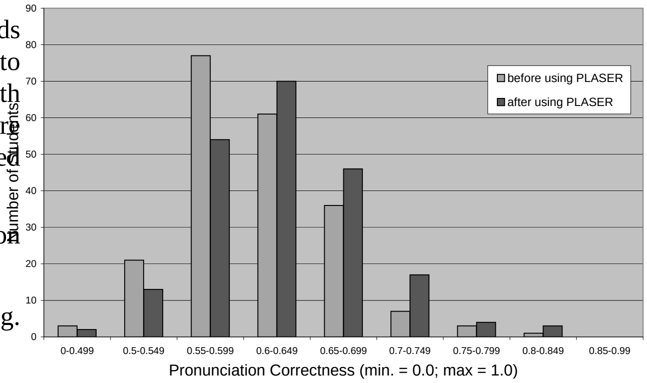


Figure 3: Results of the pronunciation evaluation test

6.1 Quantitative Evaluation

A pronunciation test consisting of speaking 60 words was conducted once before a student even started to use PLASER and once after they finished the 3-month PLASER trial. The recordings from 210 students were successfully collected. Recordings were not obtained from the rest of students for various reasons:

- Some schools did not have time to do the evaluation recordings due to schedule problems.
- Some recordings were poorly administered; e.g. parts of utterances were missing in the files.
- Some schools accidently erased or lost the recorded speech files in their computers.

At the end, recordings from 210 students were found to be good enough for evaluation. Their recordings were transcribed and compared with the correct transcriptions to find their pronunciation accuracies. The two histograms in Fig. 3 summarize their pronunciation accuracies of the 60 words before and after they practiced with PLASER. Here are some detailed statistics:

- 73% of the students had their pronunciation accuracy improved by an average of (absolute) 4.53%.
- The remaining 27% of the students got worse for unknown reasons by an average of 2.68%.
- Collectively we observe an obvious improvement: the mean accuracy after the use of PLASER is greater than that before using PLASER, and the mean difference is statistically significant at the 99% confidence level.

6.2 Qualitative Evaluation

In addition, a questionnaire survey was conducted to get comments and suggestions from teachers and students after they finished the study. Some figures are worth mentioning:

- 77% of the students believed that their pronunciation skill was improved after using PLASER while 91% of school teachers believed their students' pronunciation had improved.
- 77% of the students like to use PLASER to learn English pronunciation.
- 53% of the students preferred using PLASER to the traditional classroom teaching method to learn English pronunciation while 73% of the teachers would prefer their students using PLASER to self-learn pronunciation.
- All teachers would recommend their students to use PLASER to learn English pronunciation.

7 Discussion & Future Works

More work is being planned to further improve PLASER's performance. Robustness is the key problem. In the school environment, one simply cannot expect the students to use learning tools quietly. In addition, about forty students use a language laboratory at the same time. Since the headset microphones available in all schools are not uni-directional, recordings from neighboring students are picked up on top of the user's. This kind of "babble noise" hurts PLASER's performance to a great extent: not only does it affect the accuracy of our phoneme recognizer, various thresholds used in our confidence-based scoring and noise rejection are affected too. Various well-known robust techniques such as spectral subtraction (Boll, 1979), MLLR adaptation (Leggetter and Woodland, 1995), parallel model combination (Gales and Young, 1996), and stochastic matching (Sankar and Lee, 1996), etc. are being investigated.

To further improve phoneme discrimination, we are trying to build statistical models to test the presence of articulatory features in each phoneme (Leung and Siu, 2003). The outcome of the test will be a posterior probability of an articulatory feature which will then be combined with the score from the acoustic models with the hope to give even better accuracy.

Finally, the recognizer has to be optimized for the slower machines used in many local schools.

8 Acknowledgements

This work is supported by the Hong Kong Quality Education Fund under the grant number QEF99/00.EG01.

References

S.F. Boll. 1979. Suppression of Acoustic Noise in Speech Using Spectral Subtraction. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 24:113–120.

W. Chou. 2000. Discriminant-Function-Based Minimum Recognition Error Rate Pattern-Recognition Approach to Speech Recognition. *Proceedings of the IEEE*, 88(8):1201–1223, August.

H. Franco, L. Neumeyer, V. Digalakis, and O. Ronen. 2000. Combination of Machine Scores for Automatic Grading of Pronunciation Quality. *Speech Communications*, 30(2–3):121–130, Feb.

M.J.F. Gales and S.J. Young. 1996. Robust Continuous Speech Recognition Using Parallel Model Combination. *IEEE Transactions on Speech and Audio Processing*, 4(5):352–359, September.

Simon Ho and Brian Mak. 2003. English Pronunciation Evaluation as a Bi-threshold Phoneme Verification Problem. In *Proceedings of the European Conference on Speech Communication and Technology*, (submitted).

B.H. Juang and S. Katagiri. 1992. Discriminative Training for Minimum Error Classification. *IEEE Transaction on Signal Processing*, 40(12):3043–3054, Dec.

G. Kawai and K. Hirose. 2000. Teaching the Pronunciation of Japanese Double-mora Phonemes using Speech Recognition Technology. *Speech Communications*, 30(2–3):83–93, Feb.

C.J. Leggetter and P.C. Woodland. 1995. Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Journal of Computer Speech and Language*, 9(2):171–185, April.

K. Y. Leung and M. H. Siu. 2003. Phone Level Confidence Measure Using Articulatory Features. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*.

J. Mostow, S. Roth, A. G. Hauptmann, and M. Kane. 1994. A Prototype Reading Coach that Listens. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94)*, American Association for Artificial Intelligence, pages 785–792.

L. Neumeyer, H. Franco, V. Digalakis, and M. Weintraub. 2000. Automatic Scoring of Pronunciation Quality. *Speech Communications*, 30(2–3):83–93, Feb.

A. Sankar and C.H. Lee. 1996. A Maximum-Likelihood Approach to Stochastic Matching for Robust Speech Recognition. *IEEE Transactions on Speech and Audio Processing*, 4(3):190–202.

S.M. Witt and S.J. Young. 2000. Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning. *Speech Communications*, 30(2–3):95–108, Feb.