Comparing classifiers for pronunciation error detection

Helmer Strik [1], *Khiet Truong* [2], *Febe de Wet* [3], *Catia Cucchiarini* [1]

[1] CLST, Department of Linguistics, Radboud University, Nijmegen, The Netherlands
[2] TNO Human Factors, Soesterberg, The Netherlands
[3] SU-CLaST, Stellenbosch University, South-Africa
[h.strik|c.cucchiarini]@let.ru.nl, khiet.truong@tno.nl, fdw@sun.ac.za

Abstract

Providing feedback on pronunciation errors in computer assisted language learning systems requires that pronunciation errors be detected automatically. In the present study we compare four types of classifiers that can be used for this purpose: two acoustic-phonetic classifiers (one of which employs linear-discriminant analysis (LDA)), a classifier based on cepstral coefficients in combination with LDA, and one based on confidence measures (the so-called Goodness Of Pronunciation scores). The best results were obtained for the two LDA classifiers which produced accuracy levels of about 85-93%.

Index Terms: Computer Assisted Pronunciation Training (CAPT), pronunciation error detection, acoustic-phonetic classifiers, Goodness Of Pronunciation (GOP)

1. Introduction

Computer Assisted Language Learning (CALL) applications, and, more specifically, Computer Assisted Pronunciation Training (CAPT) applications that make use of automatic speech recognition (ASR) have received considerable attention in recent years. Most of the literature on pronunciation assessment has focused on pronunciation grading (or scoring), while less attention has been paid to error detection (or localization). Pronunciation grading usually refers to a procedure used to calculate a global pronunciation score at the speaker or utterance level, which, for that matter, could also be a weighted average of local, phoneme scores. Error detection, on the other hand, requires calculating a score at a local (e.g. phoneme) level, for each individual realization of a given phone. Although this explanation might suggest that error detection is simply a specific sub-task of pronunciation grading, in fact these are two different tasks, with different goals and different outcomes. The distinction between pronunciation scoring and error detection becomes clear when we consider the specific goals for which they are employed. Pronunciation scoring is typically used in pronunciation testing applications to calculate global scores (whether or not obtained by averaging local scores) to provide an indication of the candidate's proficiency. Such global scores are usually not informative enough for applications like pronunciation training where students usually prefer to have more specific information on the nature of their pronunciation mistakes. Therefore, in pronunciation training, information should at least be provided at phoneme level for individual realizations of the various phones, so that learners can focus their attention on the most problematic sounds.

Error detection requires a higher level of detail than pronunciation grading, which is generally based on a number of speech characteristics such as the temporal features, speech rate, articulation rate, and segment duration, which can be calculated automatically relatively easily [1] and which are measured over longer stretches of speech than the point measurements that are required for error detection. Consequently, such temporal measures are more reliable and yield stronger correlations with human judgements of pronunciation quality [2].

Various approaches to error detection can be found in the literature. The best known example is the Goodness Of Pronunciation (GOP) algorithm developed by Witt [3, 4], which has also been adopted by other authors [5, 6]. Recently, experiments have also been carried out in which classifiers using specific acoustic features, different classification methods such as Linear Discriminant Analysis (LDA) and Decision Trees and phonological features have been used [7, 8, 9, 10]. However, approaches like [10] seem more suitable for pronunciation scoring rather than for error detection, because they do not address individual realizations and do not report performance results for individual occurrences of speech sounds, but only give a rough indication of which sounds appear to be problematic for different groups of speakers.

In Truong et al. [8] we found that LDA classifiers trained on a relatively small number of phone-specific, acoustic-phonetic features (LDA-APF) manage to discriminate between voiceless fricatives and plosives in non-native Dutch and achieve 87-95% classification accuracy. In addition, the performance of LDA-APF was better than that obtained by applying a method by Weigelt et al. [11] that aimed at discrimination between voiceless fricatives and voiceless plosives. In this paper we extend the research described in [8] by studying additional approaches that make use of different input features and different methods. Specifically, we will compare LDA-APF with GOP, because this is one of the most well-known procedures, and, for a full appreciation of the effect of features (APF versus Mel Frequency Cepstrum Coefficients (MFCC)) and method (weighted versus unweighted), we will also compare LDA-APF and GOP to LDA-MFCC.

In developing our training system for Dutch pronunciation, Dutch-CAPT, we have identified 11 problematic sounds [12] on which feedback should be provided. In the current paper we focus on the discrimination of the Dutch velar fricative /x/ versus the velar plosive /k/, since the substitution of /x/ with /k/ is a typical pronunciation error in Dutch as a second language (L2). We have developed and tested four classifiers to discriminate /x/ from /k/. They are described in section 2.2. The material used to train and test these classifiers is presented in section 2.1, and the results in section 3. We end with a discussion (section 4) and conclusions (section 5).

2. Method and material

2.1. Material

For training, native speech from the Polyphone database [13] was used, consisting of read sentences, sampled at 8 kHz (telephone speech). For testing, two different sets were used: [A] native speech from the Polyphone database, and [B] non-native speech from the DL2N1 corpus (Dutch as Second Language, Nijmegen corpus 1). The DL2N1 corpus contains Dutch phonetically rich sentences that were read over the phone by 60 non-native speakers [1]. Therefore, in test condition B, there is a mismatch between training (native speech) and testing (non-native speech from another corpus).

The phonemes /x/ and /k/ were automatically extracted on the basis of time-aligned segmentations obtained with an automatic speech recognizer. The same automatic segmentation was used in all four classifiers. The number of tokens used to train and test classifiers are shown in Table 1.

Table 1. Number of tokens used for training and testing the classifiers

	Training		**Testing**			
	Native		**Condition A Native**		**Condition B Non-native**	
	M	**F**	**M**	**F**	**M**	**F**
/x/	1000	1000	2348	2279	155	230
/k/	1000[i]	1000[i]	1892[ii]	1975[ii]	162[ii]	249[ii]

2.2. Method

Below the four types of classifiers are presented. In Truong et al. [8] we already showed that the results for LDA-APF are better than those for Weigelt's method. Here we will focus on comparing LDA-APF with GOP and LDA-MFCC.

Results are presented in terms of Scoring Accuracy (SA), which is the percentage of Correct Acceptances (CA) and Correct Rejections (CR) divided by the total number of tokens (N): SA = 100% * (CA+CR) / N.

The optimization criterion used for all four classifiers was the same: maximize SA for a given maximum level of False Acceptances (10% in our case).

2.2.1. Method 1: GOP

Method 1 uses an ASR-based confidence measure, the Goodness Of Pronunciation (GOP) score [3,4]. Gender-dependent monophone HMMs were trained on 15,000 (7,500 male and 7,500 female) phonetically rich sentences from the Polyphone corpus [13]. The sentences were chosen such that the training material included at least 1,000 tokens for each phone. Twelve MFCCs, energy, and their first and second order derivatives were used. Cepstral mean normalization was implemented at utterance level to compensate for the effect of different channel properties on the data.

The GOP score for each phone corresponds to the frame-normalized ratio between the log-likelihood score of a forced and free phone recognition. If the GOP score of a specific phone falls below a certain threshold, the pronunciation of this specific phone is considered correct. As in [3], thresholds per phone were obtained by using native speech material in which errors had been artificially introduced. In our approach we introduced errors which are similar to errors found in nonnative speech (such as substitution of /k/ by /x/).

2.2.2. Method 2: Weigelt

Method 2 employs an acoustic-phonetic approach, which is based on an algorithm developed by Weigelt and colleagues [11] to discriminate between voiceless plosives and fricatives. We adopted this algorithm in our study to discriminate the voiceless velar fricative /x/ from the voiceless velar plosive /k/ [8, 14]. Weigelt's algorithm is based on three measures that can be obtained relatively easily and quickly: log root-mean-square (rms) energy, the derivative of log rms energy (the so-called Rate Of Rise (ROR)), and zero-crossing rate. Since the release of the burst of the plosive causes an abrupt rise in amplitude, the ROR values of plosives are usually much higher than those of fricatives (see Figure 1). Consequently, the magnitude of the highest peak in the ROR contour can be used to discriminate plosives from fricatives. An ROR threshold can be set to classify sounds that have an ROR peak value above this threshold as plosives, and those sounds that are characterized by an ROR peak value under this threshold, as fricatives. In addition, some criteria (mainly based on zero-crossing rate and energy) are used to discard spurious ROR peaks that are related to other speech/non-speech sounds. These criteria and the ROR thresholds were optimized for the material of the current experiment.

2.2.3. Method 3: LDA-APF

The third method uses specific (selectively chosen) acoustic-phonetic features that are potentially discriminative in a linear discriminant analysis (LDA) [8, 14]. We use ROR and log rms energy, the main features in Weigelt's algorithm, as discriminative features in LDA to discriminate /x/ from /k/. The magnitude of the highest ROR peak is used, duration, and four rms energy measurements that are made around the ROR peak at 5 ms before (i1) and at 5 ms (i2), 10 ms (i3) and 20 ms (i4) after the peak (see Figure 1). These features were extracted with Praat [15].

2.2.4. Method 4: LDA-MFCC

In method 1 GOP-scores are based on Mel-Frequency Cepstrum Coefficients (MFCCs), which are commonly employed in ASR systems, and in method 3 acoustic-phonetic features are used in combination with LDA. As an intermediate, MFCCs are used in combination with LDA in method 4. Twelve MFCCs and one energy feature are measured at the same moments that i1, i3 and i4 were extracted in method LDA-APF, making a total of 39 features.

---

[i] For the GOP method, these numbers represent the number of tokens used to determine the GOP thresholds.

[ii] For the GOP method, these numbers represent the number of tokens in the transcription where the symbol /k/ is substituted with /x/ (in order to create artificial errors).

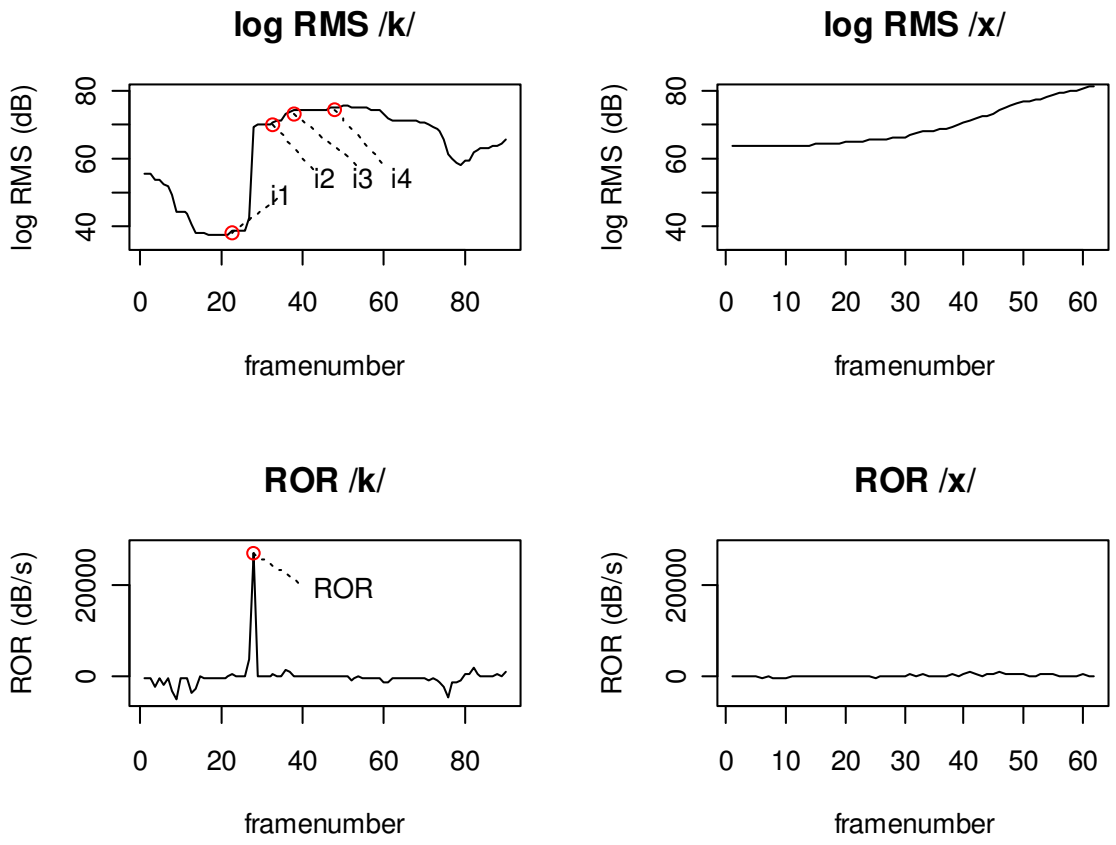

Figure 1. Log RMS (top) and ROR (bottom) contours of /k/ (left) and /x/ (right).

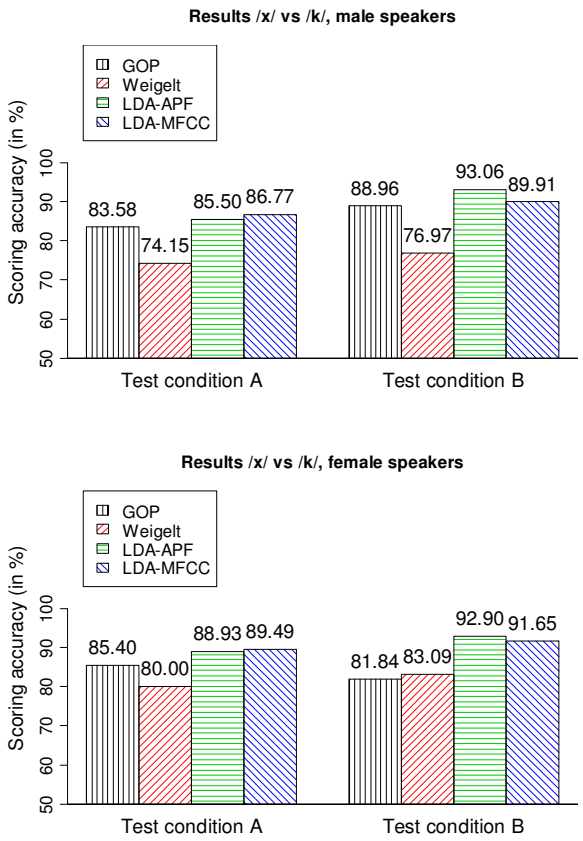

Figure 2. Scoring Accuracy (SA) values.

3. Results

Figure 2 shows Scoring Accuracy (SA) values for the 16 different combinations of

- two types of speakers: male (top) and female (bottom) speakers,
- two conditions: A (left) and B (right), and
- four classifiers, from left to right: GOP, Weigelt, LDA-APF, and LDA-MFCC.

The SA scores in Figure 2 are quite high. In all four cases (male and female, test condition A & B) the classifiers can be ordered according to decreasing SA in the following way: two LDA methods, GOP, and Weigelt. The scores for the two LDA methods are similar. In condition B (mismatch: trained on native speech, tested on non-native speech of another corpus) the scores for LDA-APF are somewhat higher than those of LDA-MFCC, while in condition A (no mismatch) it is the other way around. This would seem to indicate that LDA-APF is more robust against this 'mismatch'.

4. Discussion

We have trained classifiers for various sounds that are problematic for foreigners that learn Dutch [12,16]. Here we focus on the discrimination of /x/ vs. /k/. Results for other sounds can be found in [14]. In [8] we presented results of the comparison between Weigelt (method 2) and LDA-APF (method 3). In the present paper we present results of the comparison of these types of classifiers with two other types of classifiers: GOP scores (method 1) which have earlier been used for pronunciation error detection, and MFCCs in combination with LDA (method 4).

Ideally, one would like to train and test the classifiers using non-native pronunciation errors, since the ultimate goal

is to detect these pronunciation errors. In practice, non-native data is usually insufficient for training in general, let alone for training specific classifiers for the different non-native pronunciation errors. For this reason we decided to study the performance of the various classifiers by trying to detect non-native pronunciation errors for which the incorrect realization corresponds to another phoneme in the L2. For Dutch as L2, this is the case for the contrast between /x/ (target sound) and /k/ (incorrect realization), but also for a number of vowel errors, such as /A/-/a/, /y/-/u/ and /Y/-/u/ [12,14,16]. For these types of errors, the correct native realizations /k/, /a/, and /u/ can be used to train thresholds for detecting the non-native incorrect realizations. To further explore the performance of the classifiers and to see how they can cope with data sparseness, we also examined cases in which classifiers trained on native speech were used to detect errors in non-native speech.

The two LDA methods yielded the best performance scores followed by GOP and Weigelt. In Linear Discriminant Analysis (LDA), weights are assigned to each feature in order to find the linear combination of features which best separates the classes, while in the two other classifiers (that do not use LDA) all criteria have the same weights. For instance, in the LDA-MFCC classifier the largest weights are those of the energy features; LDA thus is capable of selecting those features that are most relevant. Apparently, this is an important advantage of the LDA-based classifiers compared to the other classifiers.

The results for the two classifiers in which LDA is used are similar. In condition B (mismatch between training and test) the results for LDA-APF were better than those of LDA-MFCC, while in condition A (no mismatch) it was the other way around. Note that in condition B the test data were taken from a different corpus, and although this corpus also contains telephone speech, the (acoustic) properties of the signals can be slightly different. Since the APF features are more specific for a given speech sound, while the MFCC features are more general in nature, it is to be expected that when there is larger mismatch between training and test data/conditions, the APF features should perform better. Our results for this limited amount of material and limited amount of mismatch seem to support this explanation. Another aspect that should be considered is the number of features employed in the two approaches. LDA-APF requires fewer features than LDA-MFCC. Additional advantages of APF features are that they are easier to interpret (compared to MFCCs), and that they can be useful for both learner (to provide meaningful feedback) and teacher (to make clear what the problematic pronunciation aspects are). On the other hand, MFCCs are already available in the ASR system and GOP scores can easily be obtained for all phones using similar procedures, while APFs have to be calculated specifically for the purpose of error detection and specific features have to be derived to train specific classifiers for every error. What is needed is a generic method to obtain acoustic-phonetic classifiers for different types of errors, and for different combinations of sounds. The best solution probably is to use both GOP and APFs in combination with LDA.

5. Conclusions

The highest scoring accuracy results were obtained for the two LDA methods, followed by GOP and Weigelt. Results for LDA-APF and LDA-MFCC are similar. Advantages of LDA-APF are that it seems to be more robust for training-test mismatches, and that fewer features are used; a disadvantage of LDA-APF is that a new classifier has to be developed for every pronunciation error.

6. References

[1] Cucchiarini, C., Strik, H. and Boves, L., “Quantitative assessment of second language learners’ fluency by means of automatic speech recognition technology”, Journal of the Acoustical Society of America 107, 989-999, 2000.

[2] Kim, Y., Franco, H. and Neumeyer, L., “Automatic pronunciation scoring of specific phone segments for language instruction”, Proceedings of Eurospeech, 645-648, 1997.

[3] Witt, S.M., Use of speech recognition in Computer-assisted Language Learning, PhD thesis, Department of Engineering, University of Cambridge, 1999.

[4] Witt, S.M. and Young, S., “Phone-level Pronunciation Scoring and Assessment for Interactive Language Learning", Speech Communication 30, 95-108, 2000.

[5] Mak, B.S., Ng, M., Tam, Y-C., Chan, Y-C., Chan, K-W., Leung, K.Y., Ho, S., Chong, F.H., Wong, J., and Lo, J., "PLASER: Pronunciation Learning via Automatic Speech Recognition", Proceedings of HLT-NAACL, 23-29, 2003.

[6] Neri, A., Cucchiarini, C. and Strik, H., "ASR corrective feedback on pronunciation: Does it really work?", Proceedings of Interspeech, 1982-1985, 2006.

[7] Tsubota, Y., Kawahara, T. and Dantsuji, M., "Recognition and verification of English by Japanese students for computer-assisted language learning system", Proceedings of Interspeech, 1205-1208, 2002.

[8] Truong, K., Neri, A., De Wet, F., Cucchiarini, C., and Strik, H., "Automatic detection of frequent pronunciation errors made by L2-learners", Proceedings of Interspeech, 1345-1348, 2005.

[9] Ito, A., Lim, Y-L., Suzuki, M., and Makino, S., "Pronunciation Error Detection Method based on Error Rule Clustering using a Decision Tree", Proceedings of Interspeech, 173-176, 2005.

[10] Stouten, F. and Martens, J.-P., "On The Use of Phonological Features for Pronunciation Scoring", Proceedings of ICASSP, 329-332, 2006.

[11] Weigelt, L.F., Sadoff, S.J. and Miller, J.D., "The plosive/fricative distinction: The voiceless case”, Journal of the Acoustical Society of America 87, 2729-2737, 1990.

[12] Neri, A., Cucchiarini, C. and Strik, H., "Segmental errors in Dutch as a second language: How to establish priorities for CAPT", Proceedings of the InSTIL/ICALL Symposium, 13-16, 2004.

[13] Damhuis, M., Boogaart, T., In 't Veld, C., Versteijlen, M., Schelvis, W., Bos, L. and Boves, L., "Creation and analysis of the Dutch Polyphone corpus", Proceedings of Interspeech, 1803-1806, 1994.

[14] Truong, K.P., Automatic pronunciation error detection in Dutch as a second language: an acoustic-phonetic approach, Master Thesis, Utrecht University, 2004.

[15] Boersma, P., "Praat: a system for doing phonetics by computer", Glot International 5:9/10, 341-345, 2001.

[16] Neri, A., Cucchiarini, C., Strik, H., “Selecting segmental errors in L2 Dutch for optimal pronunciation training”, IRAL - International Review of Applied Linguistics in Language Teaching, 44, pp. 357–404, 2006.