Comparing different approaches for automatic pronunciation error detection

Helmer Strik [a,*], Khiet Truong [b], Febe de Wet [c], Catia Cucchiarini [a]

[a] *Centre for Language and Speech Technology, Radboud University Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands*
[b] *Human Media Interaction Group, Dept. of Computer Science, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands*
[c] *Centre for Language and Speech Technology, Stellenbosch University, Private Bag X1, Matieland, 7602, South Africa*

Abstract

One of the biggest challenges in designing computer assisted language learning (CALL) applications that provide automatic feedback on pronunciation errors consists in reliably detecting the pronunciation errors at such a detailed level that the information provided can be useful to learners. In our research we investigate pronunciation errors frequently made by foreigners learning Dutch as a second language. In the present paper we focus on the velar fricative /x/ and the velar plosive /k/. We compare four types of classifiers that can be used to detect erroneous pronunciations of these phones: 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 score). The best results were obtained for the two LDA classifiers which produced accuracy levels of about 85–93%.
© 2009 Elsevier B.V. All rights reserved.

1. Introduction

Computer assisted language learning (CALL) applications and, more specifically, computer assisted pronunciation training (CAPT) applications for language learning that make use of automatic speech recognition (ASR) have received considerable attention in recent years. Most of the literature on pronunciation assessment in language learning 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,

* Corresponding author. Tel.: +31 (0) 24 361 61 04; fax: +31 (0) 24 361 29 07.
  E-mail addresses: h.strik@let.ru.nl (H. Strik), truongkp@ewi.utwen-te.nl (K. Truong), fdw@sun.ac.za (F. de Wet), c.cucchiarini@let.ru.nl (C. Cucchiarini).

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.

Automatic error detection at the phoneme level is often done by calculating so-called confidence measures (CMs) that can be derived from an ASR system based on Hidden Markov Models (HMM). Various types of CMs have been studied (e.g., Neumeyer et al., 2000). One of the best known examples is probably the Goodness Of Pronunciation (GOP) algorithm (Witt, 1999; Witt and Young, 2000), which has also been adopted by other authors (e.g., Mak et al., 2003; Neri et al., 2006a). CMs indicate how certain the recognizer is that a given target sound was pronounced correctly: the lower the confidence, the higher the chance that the sound (or utterance) was mispronounced. These CMs have the advantage that they can be obtained fairly easily with an ASR system, and they can be calculated in similar ways for all speech sounds.

To be of value in computer assisted pronunciation training systems, this type of evaluation should reflect human judgments of the same speech. However, correlations between CMs and human judgments are usually lower when they are computed for individual sounds than when they are computed over relatively long stretches of speech, e.g. a few utterances (e.g., Neumeyer et al., 2000). The smaller correlations observed at phoneme level may partly be explained by the lack of specificity that characterizes CMs, which are calculated using similar procedures for the same set of features for all the sounds, while this might not be the optimal approach to detect all pronunciation errors.

In a study of pronunciation errors in Dutch as a second language (L2), we found that learners of Dutch have problems with vowel length, and fricatives vs. plosives (Neri et al., 2004, 2006b). This study was aimed at selecting the pronunciation errors that should be addressed in an ASR-based pronunciation training program for Dutch L2 called *Dutch-CAPT*. The aim of this training program was to help correct salient pronunciation problems that might hamper communication. The program addressed pronunciation errors that were selected on the basis of the following six criteria:

1. Common across speakers of various L1's.
2. Perceptually salient.
3. Potentially hampering communication.
4. Frequent.
5. Persistent.
6. Suitable for robust automatic detection.

Pronunciation errors that result in the realization of a Dutch phoneme different from the intended one and which lead to a different word being uttered are likely to hamper communication as they might result in a different message being understood. This might be the case when a long vowel is realized instead of a short one, thus producing the word /maːn/ (moon) instead of /mɑn/ (man), or, analogously, when a velar plosive is realized instead of a velar fricative, for instance in words such as /xaːs/ (gauze) and /xɑs/ (gas), thus leading to the realization of the words /kaːs/ (cheese) and /kɑs/ (cash desk), respectively.

It is well known that HMMs are not very powerful devices to discriminate between sounds that are spectrally similar and differ mainly in duration; and HMMs are also not very well suited to discriminate between fricatives and plosives, since the difference between plosives and fricatives is coded in subtle differences in the amplitude envelope of the sound (Weigelt et al., 1990), a feature that does not figure prominently in HMMs. One could think that by focusing more on these specific acoustic–phonetic details it might be possible to train classifiers that better discriminate mispronounced sounds from the correct, target sounds. This is where algorithms aimed at discrimination between specific phonetic classes, e.g. fricatives vs. plosives (Weigelt et al., 1990), can come in handy.

Within the framework of our research on automatic pronunciation training for Dutch as a second language (DL2), we decided to explore the potential of specific acoustic phonetic feature analysis for automatic pronunciation error detection. We studied classifiers to discriminate long vs. short vowels, and fricatives vs. plosives (Truong, 2004). Here we focus on the discrimination of a typical pronunciation error in DL2 which is the substitution of the Dutch velar fricative /x/ with the velar plosive /k/. To select the specific acoustic phonetic features needed for this discrimination task, we draw on a previous study by Weigelt et al. (1990) which describes an algorithm that was able to discriminate voiceless fricatives from voiceless plosives with high accuracy.

We implemented and tested four different classifiers: (1) the GOP method; (2) a method based on (Weigelt et al., 1990); (3) the LDA-APF method in which Acoustic–Phonetic Features (APF) based on (Weigelt et al., 1990) are combined with Linear Discriminant Analysis (LDA); and (4) the LDA-MFCC method in which LDA is combined with Mel-Frequency Cepstrum Coefficients (MFCCs). The MFCCs are standard features which are generally used in ASR systems, e.g. for computing GOP-scores. Methods 1 and 4 are thus based on MFCCs, while methods 2 and 3 are based on acoustic–phonetic features; and methods 3 and 4 use LDA, while this is not the case for methods 1 and 2.

These four classifiers are described in Section 2.3, which is the main part of Section 2. The classifiers were compared in a series of experiments of which the results are presented in Section 3. These results open up new avenues of research and new possibilities of applications of these types of algorithms to automatic error detection in computer assisted pronunciation training. These and other issues are touched upon in Section 4.

2. Material and method

2.1. Material

All four classifiers used the same material for training and testing (see Table 1). We trained with native speech data taken from the Polyphone database (Den Os et al., 1995): read sentences, sampled at 8 kHz (telephone speech). For testing, we used two different conditions: (a) native speech from the same Polyphone database, and (b) non-native speech from the DL2N1 corpus (Dutch as Second Language, Nijmegen corpus 1). The DL2N1 corpus was collected in a previous study (Cucchiarini et al., 2000). In this corpus, 60 non-native speakers called from their home and read 10 Dutch phonetically rich sentences over the telephone. The subjects were sufficiently varied with respect to their mother tongue and proficiency level in Dutch. The speech (sampled at 8 kHz) was orthographically transcribed and automatically segmented by a speech recognizer.

The speech of 31 low-proficient non-native speakers of Dutch was annotated on pronunciation errors by a trained linguist. Because annotations of pronunciation errors are labour-intensive and costly, we decided to have one expert carry out the task. Since it is well known that phonetic transcriptions may contain an element of subjectivity (Cucchiarini, 1996), we tried to minimize the problem by choosing a qualified annotator who had previously carried out similar tasks following a consensus procedure with other experts. To determine whether the degree of subjectivity in our annotations was acceptable, we asked this annotator and five other Dutch experts to transcribe errors for a subset of the speech fragments and then compared the annotations pairwise using the transcription alignment program described in (Cucchiarini, 1996). The degree of agreement between our annotator and the five experts was generally very high: $M = 89.3\%$, $SD = 2.57$. More details on the procedure adopted to check the reliability of the annotations are provided in (Neri et al., 2006b).

The phonemes /x/ and /k/ were automatically extracted from the signals based on a segmentation that was obtained by carrying out a forced alignment with the automatic speech recognition (ASR) system described in Section 2.2. The same automatic segmentation was used for all experiments.

2.2. Automatic speech recognition system

The ASR system consists of 37 context independent monophone hidden Markov phone models, and was imple-

mented in the Hidden Markov Models Toolkit (HTK; Young et al., 2000). The phone set includes a general speech model to account for unintelligible speech as well as a silence and a short pause model. Except for the short pause model, each hidden Markov model (HMM) has three states and 32 Gaussian mixtures per state. The single state of the short pause model is tied to the central state of the silence model. The HMMs are gender-dependent and were trained on 15,000 (7500 male and 7500 female) phonetically rich sentences from the Polyphone corpus (Den Os et al., 1995). The sentences were chosen such that the training material included at least 1000 tokens for each phone.

2.3. Classifiers

In this section we describe four different methods to perform pronunciation error detection.

2.3.1. Method GOP: Goodness-Of-Pronunciation

The CMs that are derived from the ASR system, are obtained by a method suggested by Witt and Young (2000). Gender-dependent HMMs are trained on native training data and are used to perform free phone and forced recognition for each utterance. The Goodness-of-Pronunciation (GOP) score is subsequently derived at phone level. The GOP score for each phone corresponds to the frame normalized ratio between the log-likelihood score of its forced and free phone recognition. If the GOP score of a specific phone falls below a certain threshold, the pronunciation of the phone is accepted as a correct instance of the phone and vice versa. The GOP score is calculated in the same way for each phone, without taking specific properties of the phone into account.

According to the method proposed in (Witt and Young, 2000), the mean and variance of the GOP scores obtained from the training data can be used to choose an appropriate threshold for each phone. However, the resulting range of threshold values did not yield meaningful thresholds for our data. We therefore carried out an exhaustive search for values in the range between 0 and 10. We used an optimization criterion similar to that adopted by Witt (1999), maximizing the Scoring Accuracy (Correct Acceptance + Correct Rejection) while trying to keep the percentage of False Rejections below 10%.

Ideally, one should have enough non-native data available to derive the thresholds from real target data. Unfortunately, we did not have such a corpus available and therefore had to use artificial data that were generated on the basis of information obtained from the studies on pro-

Table 1
Number of tokens used to train and test the models.

	Training (native, Polyphone)		Testing condition A (native, Polyphone)		Testing condition B (non-native, DL2N1)	
	Male	Female	Male	Female	Male	Female
/x/	1000	1000	2348	2279	155	230
/k/	1000	1000	1892	1975	162	249

nunciation errors made by L2 learners of Dutch reported on in (Neri et al., 2006b). These studies indicated that, for instance, /x/, /ɑ/, /y/ and /ɪ/ are often replaced by: /k/, /a/, /u/ and /i/, respectively. The artificial data were created by introducing errors into the pronunciation dictionary: for the present study incorrect pronunciations were "created" by changing the transcriptions of the phonemes /x/ and /k/.

2.3.2. Method Weigelt: decision-tree

In (Weigelt et al., 1990), an algorithm is described that successfully discriminates voiceless fricatives from voiceless plosives. We adopted this algorithm in our study to discriminate the voiceless velar fricative /x/ from the voiceless velar plosive /k/. Since the gross spectral shapes of plosives and fricatives are similar, several other cues to discriminate between these classes have been investigated, most of them based on the amplitude envelope. Weigelt's algorithm is based on three measures that can be obtained relatively easily and quickly: log rms (root-mean-square) energy, the first-order derivative of log rms energy, and zero-crossing rate. The main feature is the so-called Rate Of Rise (ROR) which is calculated by determining the derivative of the logarithm of rms energy. A 24 ms window n is shifted every 1 ms over the acoustic signal and for each window n, the energy is measured by computing the logarithm of the rms over window n:

$$E_n = 20 \times \log 10\left(\frac{rms_n}{0.00002}\right)$$

ROR, the energy derivative, is then computed as:

$$ROR_n = \frac{(E_n - E_{n-1})}{\Delta t}$$

where Δt is the time step in which the window n is shifted, in our case 1 ms.

The zero-crossing rate is defined as the number of times the signal crosses the axis. 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 Fig. 1). Consequently, the magnitude of the largest 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 (Weigelt et al., 1990), this threshold is heuristically set at 2240 dB/s. However, large peaks in the ROR contour can also be the result of other speech (e.g., vowel onset) or non-speech sounds (e.g., lip smacks). Therefore, the algorithm by Weigelt et al. (1990) uses four criteria to distinguish non-significant ROR peaks (i.e. those belonging to other speech/non-speech sounds) from significant ROR peaks (i.e. those that do belong to the release of the burst of the plosive). Starting with the largest peak, each criterion is examined: (1) for the 49-ms period following the peak, the value of E_n must never fall below the value of E at the peak,

(2) the maximum value of E_n for the following 49 ms must be at least 12 dB above the value of E at the peak, (3) the maximum zero-crossing rate over the 49-ms period after the peak must be higher than 2000 zero crossings per second, and (4) the zero-crossing rate at exactly 49 ms after the peak must be no more than 100 crossings per second below the zero-crossing rate 20 ms before the peak. The derivative of log rms energy is a natural candidate for a measure that may capture differences in abrupt and slow rises of energy. Zero-crossing rates are used to discard peaks related to other speech/non-speech sounds. The duration of an energy pulse is examined to ensure that it is large enough in both amplitude and duration to be considered a speech sound. If any of these criteria fails, then the peak is labeled as non-significant and the phoneme is classified as a voiceless fricative. If the peak is significant and its ROR value is above the predetermined threshold, then the phoneme is classified as a voiceless plosive. All criteria were set and tuned by training and testing the algorithm using all possible combinations of values for the criteria. During this tuning stage, criterion four appeared to be too strict and was therefore not used in our adapted algorithm. The algorithm was optimized by adopting a similar criterion as applied in the GOP method. All acoustic measurements were carried out automatically in the program Praat 8 (Boersma, 2001) and were based on the same automatic segmentation that was used to compute the GOP measures. Fig. 2 summarizes the adapted Weigelt binary decision-tree used in our study.

2.3.3. Method LDA-APF: Linear Discriminant Analysis with Acoustic–Phonetic Features

The relatively simple algorithm by Weigelt et al. (1990) proved to be successful in discriminating plosives from fricatives with energy as the primary acoustic–phonetic feature. However, in our adapted Weigelt's binary decision-tree, all criteria have the same weights. If any of the proposed criteria fails, then the phoneme is classified as a fricative, regardless of the importance of that criterion. In Linear Discriminant Analysis (LDA), weights are assigned to each feature in order to find the linear combination of features which best separates two or more classes. 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, irrespective of its significance, is used, including 4 rms energy measurements that are made around this peak at 5 ms before ($i1$) and at 5 ms ($i2$), 10 ms ($i3$) and 20 ms ($i4$) after the peak (see Fig. 1). In addition to these energy features, duration, either raw (dur) or normalized ($norm$-dur) is optionally added as a feature. Duration could be a discriminative cue since plosives are typically short acoustic events, while fricatives tend to be somewhat longer in duration. As non-natives tend to have lower articulation rates and longer segment durations (Cucchiarini et al., 2000) duration was normalized for articulation rate (articulation rate is defined as the number of speech sounds divided by
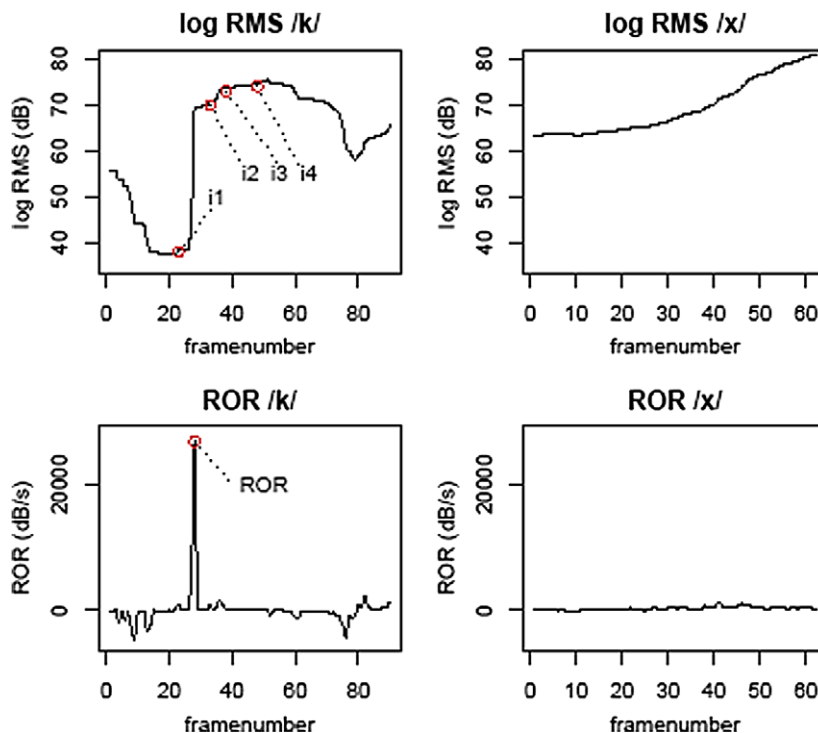
Fig. 1. Log rms (top) and ROR (bottom) contours of /k/ (left) and /x/ (right).

the total duration of the utterance without internal pauses). Duration normalization per speaker was achieved by com-

Fig. 2. Summary of decision-tree Weigelt, each criterion has to pass a certain threshold T.

puting the product of the articulation rate per speaker and segment duration:

normalized duration = articulation rate × segment duration

All acoustic measurements were made in Praat and were based on the same automatic segmentation that was used for the other methods. The LDA models were trained using the same optimization criterion as in the GOP method.

2.3.4. Method LDA-MFCC: Linear Discriminant Analysis with Mel-Frequency Cepstrum Coefficients

The CMs calculated in method GOP are based on Mel-Frequency Cepstrum Coefficients (MFCCs). It is standard practice in ASR systems to use MFCCs, together with their first and second-order derivatives (which are often referred to as delta and delta–delta, respectively). In order to enable a fairer comparison between the LDA-APF and GOP classifiers, we decided to develop another classifier, i.e. LDA-MFCC, in which LDA is combined with MFCCs. Twelve MFCCs and one energy feature are calculated at the same moments that *i1*, *i3* and *i4* were extracted in method LDA-APF, together with their first and second-order derivatives. These 39 MFCC features were used to train (again using the same optimization criterion) and test the classifiers.

2.4. Evaluation

A classification algorithm like the GOP can produce four types of outcomes: (1) correctly accepted (CA) phone
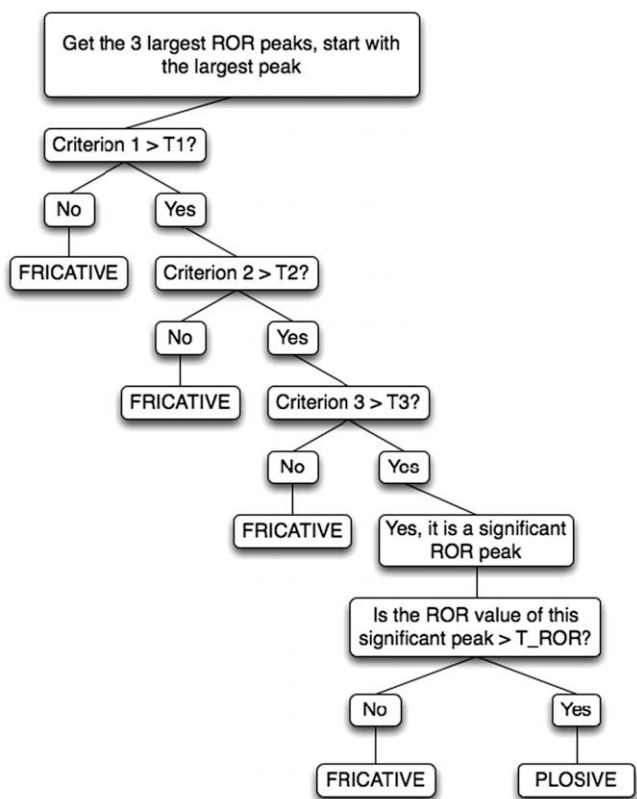
realizations, i.e. phones that were pronounced correctly and were also judged as correct; (2) correctly rejected (CR) phone realizations, i.e. phones that were pronounced incorrectly and were also judged as incorrect; (3) mispronunciations that were falsely judged as being correct (FA: False Accept) and (4) correct pronunciations that were falsely flagged as mispronunciations (FR: False Reject). To achieve optimal performance the algorithm should detect the mispronunciations and, at the same time, it should not flag as mispronunciations those realizations that were actually correct. For this reason both the number of correctly rejected (CR) and correctly accepted (CA) realizations are important in the performance calculation.

The performance levels of the classifiers are evaluated by calculating the Scoring Accuracy, which is defined as:

Scoring Accuracy

$$= 100 \times \left(\frac{\text{Correct Acceptances} + \text{Correct Rejections}}{\text{Total number of tokens}} \right)$$

3. Results

The four different classifiers, all trained on native speech, were tested on native (condition A) and non-native speech (condition B), for data of male and female speakers. This makes a total of 16 different combinations: 4 classifiers × 2 test conditions × 2 genders. The scoring accuracy (SA) values of these 16 different combinations are presented in Fig. 3. The statistical significance of the differences between the different test conditions was determined by performing a z-test for the difference between independent proportions. Bonferroni correction for multiple testing was incorporated in the test. These results are presented in Table 2.

In Fig. 3 it can be observed that all SA scores 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. However, these differences are not statistically significant. For test condition A, the SA scores are always higher for the data of the female speakers; these differences (between the male and female results) are significant for all the methods except GOP. Mixed results are obtained for condition B: higher SA scores for female data for the methods Weigelt and LDA-MFCC. Although the differences in SA are quite large, they are not significant in this case. This observation can be explained by the fact that the number of samples in the test population is much smaller in condition B than in condition A. Absolute differences in SA therefore need to be much larger in condition B in order to be significant.
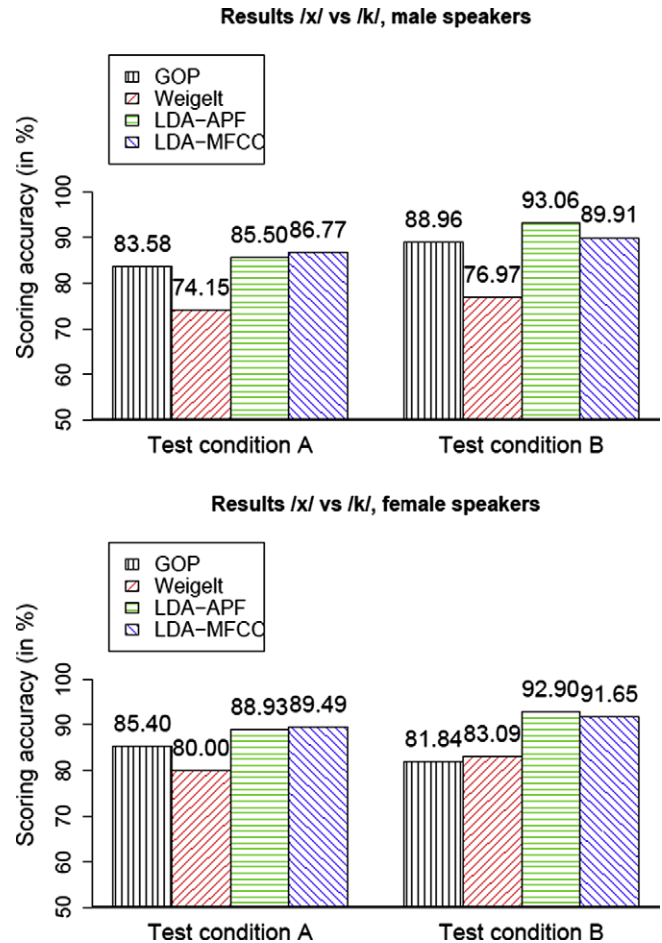

Fig. 3. Results of classification experiments, /x/ vs. /k/, for male (top) and female (bottom) speakers.

Table 2
Statistical significance of the differences observed in Fig. 3. Terminology: '=': difference is not significant, '>' or '<': significantly (p < 0.05) smaller or larger, respectively.

	GOP		LDA-APF		LDA-MFCC	
	m	f	m	f	m	f
(a) Test condition A						
Weigelt	<	<	<	<	<	<
GOP			=	<	<	<
LDA-APF					=	=
(b) Test condition B						
Weigelt	<	=	<	<	<	<
GOP			=	<	=	<
LDA-APF					=	=

4. Discussion and conclusions

4.1. Discussion

Several classifiers for automatic pronunciation error detection were developed and compared. These classifiers were developed for cases that are known to be problematic for foreigners that have to learn Dutch (Neri et al., 2004, 2006b): vowel length (e.g., /ɑ/ vs /aː/, /Y/ vs /u,y/) , and

pronunciation of fricatives such as the /x/ (which is often mispronounced as the plosive /k/). In the present study, we focus on the mispronunciation of /x/ as /k/; results for other types of pronunciation errors can be found in (Truong, 2004). Four different methods to discriminate between /x/ and /k/ were compared to each other: two methods based on commonly used MFCC features (GOP and LDA-MFCC), and two other methods based on Acoustic–Phonetic Features (Weigelt and LDA-APF).

In practice, the amount of non-native speech data, and especially the number of non-native pronunciation errors, is not sufficient for reliably training and testing classifiers. Witt proposed a method to introduce pronunciation errors in native data artificially, and used these artificial errors for developing the GOP classifier (Witt, 1999). In this study we also artificially introduced errors in native speech. These artificially introduced errors were based on errors that are really made by non-natives, e.g. the incorrect pronunciation of /x/ as /k/. The test on native speech (condition A) was carried out to examine how well the discrimination between /x/ and /k/ could be made in "normal" native speech: the outcomes could serve as an upper-limit in performance. The test on non-native speech (condition B) was carried out to examine how well this distinction was made in non-native speech and to investigate whether the thresholds obtained with artificial errors in native speech were also suitable for non-native speech. The expectation was that the discrimination between /x/ and /k/ would be more difficult to make in non-native speech since non-native speakers usually utter speech sounds with a certain foreign accent that does not sound native-like, thus making the distinction between /x/ and /k/ less clear-cut. However, the results in Fig. 3 do not support this, which suggests that, at least in this case, foreign pronunciation errors can be detected using models that are trained with artificial errors in native speech. This is a considerable advantage given that data sparseness is a common problem in research on ASR-based L2 pronunciation error detection.

The results (Fig. 3) indicate that both LDA-APF and LDA-MFCC perform better than 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 is thus 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. When we look more closely at the differences between LDA-APF and LDA-MFCC we can observe that (1) LDA-APF performs better than LDA-MFCC in the condition B, but not in the condition A, and (2) the differences in performance between LDA-APF and LDA-MFCC are rather small and not statistically significant. Although LDA-APF and LDA-MFCC appear to achieve similar performance, from an efficiency and informative point of view,

we prefer LDA-APF to LDA-MFCC: LDA-APF uses fewer features than LDA-MFCC, the features of LDA-APF are more intuitive (compared to MFCC's), and they can be useful for both learner (to provide meaningful feedback) and teacher (to make clear what the problematic pronunciation aspects are).

Furthermore, 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. Being more robust to training-test mismatches is another advantage of the APF-based classifiers.

In general, the advantage of an LDA-APF classifier is that one can focus on specific aspects of the error it has to detect, using features that are most suitable, making use of existing knowledge. However, an important disadvantage is that a specific classifier has to be developed for every pronunciation error. 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. However, the LDA-MFCC and GOP methods are also phone/error specific in that a phone-specific LDA matrix needs to be derived and a GOP threshold determined for each new pronunciation error that needs to be investigated. To increase the applicability of APF-based classifiers to other pronunciation errors, a more generic method to obtain acoustic–phonetic classifiers for these errors is needed. One of the directions that can be followed in future research is to develop training schemes that can automatically select the features that are most suitable for detecting each specific error.

4.2. Conclusions

The results of this study have shown that, for the correct classification of the velar fricative /x/ and the velar plosive /k/, both the LDA-APF and the LDA-MFCC methods yield better results than classification based on GOP scores and the decisions tree in Weigelt's method.

Although the LDA-APF method does not consistently outperform the LDA-MFCC method, it is more appealing from an efficiency and informative point of view because LDA-APF uses fewer features, the features that are used are relatively easy to extract and, in contrast to MFCCs, there is a clear link between feature properties and pronunciation errors.

References

Boersma, P., 2001. Praat, a system for doing phonetics by computer. Glot Internat. 5:9/10, 341–345.

Cucchiarini, C., 1996. Assessing transcription agreement: methodological aspects. Clin. Linguist. Phonet. 102, 131–155.

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, 989–999.

Den Os, E.A., Boogaart, T.I., Boves, L., Klabbers, E., 1995. The Dutch polyphone corpus. In: Proc. Eurospeech, pp. 825–828.

Mak, B.S., Ng, M., Tam, Y.-C., Chan, Y.-C., Chan, K.-W., Leung, K.Y., Ho, S., Chong, F.H., Wong, J., Lo, J., 2003. PLASER: pronunciation learning via automatic speech recognition. In: Proc. HLT-NAACL, pp. 23–29.

Neri, A., Cucchiarini, C., Strik, H., 2004. Segmental errors in Dutch as a second language: how to establish priorities for CAPT. In: Proc. InSTIL/ICALL Symposium, pp. 13–16.

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

Neri, A., Cucchiarini, C., Strik, H., 2006b. Selecting segmental errors in L2 Dutch for optimal pronunciation training. IRAL – Internat. Rev. Appl. Linguist. Lang. Teaching 44, 357–404.

Neumeyer, L., Franco, H., Digalakis, V., Weintraub, M., 2000. Automatic scoring of pronunciation quality. Speech Comm. 30, 83–94.

Truong, K.P., 2004. Automatic Pronunciation Error Detection in Dutch as a Second Language: An Acoustic–Phonetic Approach. MA Thesis, Utrecht University.

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

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

Witt, S.M., Young, S., 2000. Phone-level pronunciation scoring and assessment for interactive language learning. Speech Comm. 30, 95–108.

Young, S., Kershaw, D., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2000. The HTK Book Version 3.0. Cambridge University, Cambridge.