

Automatic pronunciation error detection: an acoustic-phonetic approach

Khiet Truong

Utrecht University
K.Truong@let.kun.nl

Ambra Neri

University of Nijmegen
A.Neri@let.kun.nl

Catia Cucchiarini

University of Nijmegen
C.Cucchiarini@let.kun.nl

Helmer Strik

University of Nijmegen
H.Strik@let.kun.nl

Abstract

In this paper, we present an acoustic-phonetic approach to automatic pronunciation error detection. Classifiers using techniques such as Linear Discriminant Analysis or a decision tree were developed for three sounds that are frequently pronounced incorrectly by L2-learners of Dutch: /A/, /Y/ and /x/. The acoustic properties of these pronunciation errors were examined so as to define a number of discriminative acoustic features to be used to train and test the classifiers. Experiments showed that the classifiers are able to discriminate correct sounds from incorrect sounds in both native and non-native speech, and therefore can be used to detect pronunciation errors in non-native speech.

1 Introduction

In order to help L2-learners improve their pronunciation, it is desirable to give feedback on various aspects of pronunciation, among which the phonetic quality of the speech sounds. To this end, it is necessary to detect the L2 sounds that are most problematic for L2-learners. This paper is about developing and training classifiers for L2-pronunciation error detection.

Many methods for automatic pronunciation error detection use confidence measures computed by automatic speech recognition (ASR) software. These measures have the advantage that they can be obtained fairly easily, and that they can be calculated in similar ways for all speech sounds. However, ASR confidence measures also have the disadvantage that they are not very accurate: the average human-machine correlations they yield are rather low, and, consequently, their predictive power is also rather low (see e.g. [1]). This lack of accuracy might be related to the fact that confidence scores are computed in the same way for all speech sounds, without focusing on the specific acoustic-phonetic features of individual sounds.

Given the disadvantages of methods based on confidence measures, we have been looking for alternative approaches that would yield higher detection accuracy. In this paper we report on a study in which an acoustic-phonetic approach to automatic pronunciation error detection was investigated. This approach enables us to be more specific and, probably thereby, to achieve higher error detection accuracy and higher human-machine agreement. More

specificity is achieved in two ways. First, by examining the acoustic differences between the correct sound and the mispronounced one and by using these acoustic differences to develop classifiers for each specific pronunciation error. Second, by developing gender-dependent classifiers in which each classifier is optimally adapted to a male or a female voice. Furthermore, the acoustic-phonetic approach enables us to examine the relative importance of individual acoustic features by using Linear Discriminant Analysis (LDA).

For the current study, a survey of pronunciation errors made by L2-learners of Dutch was conducted (see section 2.1.2 and [2]). This survey revealed that the sounds /A/, /Y/ and /x/ are often pronounced incorrectly by non-native speakers, irrespective of their L1. Next, acoustic differences between correct and incorrect sounds were examined, which resulted in the selection of a number of potentially discriminative features (section 2). Finally, the classifiers based on the selected acoustic features were trained and tested (section 3) to check whether they were able to discriminate between correct and incorrect sounds. This research was carried out within the framework of the MA thesis of the first author. Classifiers were developed for each pronunciation error of /A/, /Y/ and /x/. In this paper, we will focus on the /x/-classifier. Finally, a short summary of the results for /A/ and /Y/ will be given at the end of this paper.

2 Material and Method

2.1 Material

2.1.1 Corpus

We used the DL2N1 corpus (Dutch as L2, Nijmegen corpus 1) which contains speech from native and non-native speakers of Dutch. This corpus was collected in a previous study, for more details see [3]. Subjects called from their home and read aloud ten Dutch phonetically rich sentences over the telephone. Their speech was recorded by a system connected to the ISDN line and was sampled at 8 kHz. All speech was orthographically transcribed and automatically segmented by the speech recognizer (HTK) using the Viterbi algorithm.

The native part of the corpus (DL2N1-Nat) consists of speech from 4 speakers of Standard Dutch and 16 speakers of regional varieties of Dutch. The non-native part (DL2N1-NN) consists of speech from

60 non-native speakers. This non-native group is sufficiently varied with respect to L1 and proficiency level in Dutch.

For the classification experiments, all material was divided into training data (75%) and test data (25%). Furthermore, the material was divided into male speech and female speech to develop gender-dependent classifiers.

2.1.2 Material used in classification experiments

To determine the frequency of pronunciation errors, a survey was carried out on DL2N1-NN (see [2] for more details). The speech of 31 (12 male and 19 female) non-natives was annotated on segmental pronunciation errors by expert listeners. On the basis of this survey we decided to select the segmental pronunciation errors shown in Table 1 for the present study.

	Error	Mispronounced as
Most frequent for vowels	/A/	/a:/
	/Y/	/u/ or /y/
Most frequent for consonants	/x/	/k/ or /g/

Table 1. Segmental pronunciation errors addressed in this study (phonetic symbols in SAMPA notation).

Since in the non-native annotated material the number of realizations of /a:/, /u/, /y/, /k/ and /g/ that result from pronunciation errors was too low to train and test acoustic-phonetic classifiers, we decided to study how well the classifiers can discriminate /A/, /Y/, and /x/ from correct realizations of /a:/, /u/, /y/ and /k/, respectively. Thus, all classifiers investigated in this paper were trained on tokens that were considered as pronounced correctly (for numbers of tokens used for the /x/-/k/ classifier, see Table 2). We did not include the /g/, since this sound is uncommon in Dutch and therefore we did not have enough training material. Separate classifiers for the three errors were trained, i.e. one to discriminate /A/ from /a:/, one for /Y/ and /u,y/, and one for /x/ versus /k/.

	/x/		/k/	
	Training	Test	Training	Test
DL2N1-Nat Male	84	28	89	30
DL2N1-Nat Female	126	43	126	42
DL2N1-NN Male	116	39	121	41
DL2N1-NN Female	172	58	186	63

Table 2. Numbers of correctly pronounced tokens that were used to train and test the /x/-/k/ classifiers.

2.2 Method

2.2.1 Method I

In [4], a non-statistical algorithm that successfully discriminates voiceless fricatives from voiceless plosives is described. This algorithm, which can be seen as a decision tree, was adopted in our study to discriminate the voiceless velar fricative /x/ from the voiceless velar plosive /k/. The main feature used in this algorithm is ROR (Rate Of Rise), which is calculated as described below.

A window n of 24ms long is shifted every 1ms over the acoustic signal and for each window n the amplitude is measured by computing the logarithm of the Root-Mean-Square over window n :

$$E_n = 20 * \log_{10} (RMS_n / 0.00002)$$

ROR is then computed:

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

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

Since the rise of amplitude is usually (much) higher in plosives than in fricatives, the magnitude of the peaks 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 above this threshold, like plosives, and those that are characterized by an ROR peak under this threshold, like fricatives. In [4] this threshold is 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 smack). Therefore, four criteria, of which three were used in our implementation because the fourth one appeared to be too strict, were defined to distinguish non-significant ROR peaks from significant ROR peaks, starting with the highest ROR peak: 1) for the 49-ms period following the peak, the value of E must never fall below the value of E at the peak, 2) the maximum value of E for the following 49 ms must be at least 12 dB above the value of E at the peak, and 3) the maximum zero-crossing rate over the 49-ms period after the peak must be higher than 2000 zero crossings per second. If any of these criteria fails, then the peak is not significant and the consonant is classified as a fricative. If the peak is significant and its ROR value is above the predetermined threshold, then the sound is classified as a plosive. All thresholds were set and tuned heuristically (which was done in [4] as well) by training and testing the algorithm automatically with varying values for the thresholds.

2.2.2 Method II

In addition to the algorithm in [4], another method was developed that uses Linear Discriminant Analysis (LDA) to discriminate /x/ from /k/. A number of

3 Results

potential discriminative acoustic features were employed in this LDA method: amplitude, highest ROR value and duration. Duration, either raw or normalized, was chosen because fricatives are usually longer than plosives. The highest ROR peak was taken, irrespective of its significance. Duration had to be normalized for articulation rate (defined as the number of sounds divided by total duration of the utterance without internal pauses), because, as shown in [3], non-natives have lower articulation rates and longer segment durations. Duration normalization per speaker was done as follows:

$$\text{normalized duration} = \text{art.rate} \times \text{segment duration}$$

Additionally, four amplitude measurements were taken to model the amplitude contour: $i1$ at 5ms before the highest ROR peak, and $i2, i3, i4$ at 5, 10 and 20ms, respectively, after the highest ROR peak. In total, 6 features were used in the LDA method: $ROR, i1, i2, i3, i4$ and duration (either raw ‘*rawdur*’, normalized ‘*normdur*’, or not used at all ‘*nodur*’).

All acoustic measurements were based on the automatic segmentation and were carried out automatically by *Praat* (a tool for acoustic analysis, [5]).

2.2.3 Experiments A and B

Two types of experiments were carried out. In experiment A, training and test data were taken from the same corpus with the same type of speech: A.1 is trained and tested on native data and A.2 is trained and tested on non-native data (see Table 3). Experiment A was mainly carried out to test whether the methods developed were able to discriminate between /x/ and /k/, and to examine the relative importance of each feature in the LDA method.

In experiment B, training and test data were taken from two different types of speech and were applied to each other: non-native test data is applied to a natively trained classifier. Experiment B was carried out to investigate how a natively trained classifier would cope with non-native speech: what is the performance of the natively trained classifier, as compared to a non-natively trained classifier that is already adapted to non-native speech (exp. A.2)?

Exp.	Training	Test
A.1	DL2N1-Nat	DL2N1-Nat
A.2	DL2N1-NN	DL2N1-NN
B.1	DL2N1-Nat	DL2N1-NN

Table 3. Experiments with different train and test conditions.

3.1 Classification results /x/

3.1.1 Method I

The algorithm by [4] was first trained on native or non-native data to determine the values for the thresholds used in the algorithm. Many values from the original algorithm needed to be adjusted, because their criteria appeared to be too strict; for the same reason, one criterion from the original algorithm was discarded. In Table 4, the classification results obtained with this method under different training conditions are shown separately for male and female speakers. The results range from 75.0% to 91.7% correct classification: for instance, in the A.1 experiment 81.0% (male) and 75.3% (female) of all /x/ and /k/ occurrences were correctly classified.

Experiment	M	F
A.1 Training & Test = DL2N1-Nat	81.0%	75.3%
A.2 Training & Test = DL2N1-NN	80.0%	91.7%
B.1 Training = DL2N1-Nat Test = DL2N1-NN	75.0%	91.7%

Table 4. Results from Method I, adjusted algorithm from [4].

It seems that the algorithm is able to discriminate between /x/ and /k/. Furthermore, applying a natively trained classifier to non-native speech (exp. B.1) slightly reduces the performance for male speech, but not for female speech.

3.1.2 Method II

The second method uses LDA as a classification technique to discriminate /x/ from /k/. Experiments and LDA-analyses (LDA offers a number of ways of pruning away less significant features) that were carried out on all features $ROR, i1, i2, i3, i4$ and $nodur/rawdur/normdur$ made it clear that not all features were needed to discriminate between /x/ and /k/. With only two or three features ([i1 i3] or [ROR i3], with duration optionally added), classification results ranging from approximately 85% to 95% were observed (see Fig.1). The addition of duration, with somewhat better results for normalized duration, resulted in small improvements in classification accuracy in A.1 (Fig.1). In experiment A.2 (Fig.1), on the other hand, duration (either raw or normalized) did not seem to be discriminative. Apparently, the non-native speakers of DL2N1-NN do not produce a length difference between /x/ and /k/, whereas native speakers of DL2N1-Nat usually do.

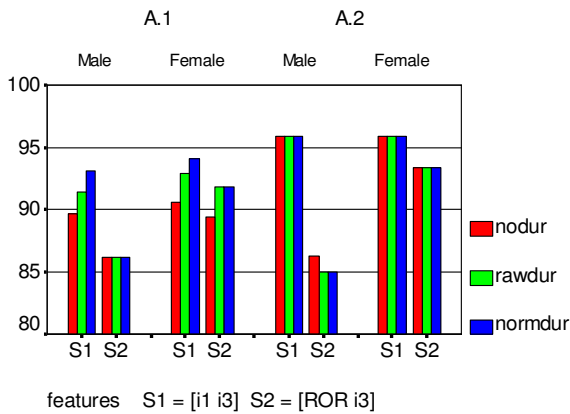


Figure 1. Correct classification %, left = A.1 right = A.2.

Furthermore, the height of the ROR peak (*ROR*), which is the main feature in method I, is less important or even superfluous in method II: the classification accuracy is higher for [i1 i3] than for [ROR i3], implying that (in combination with *i3*) *i1* is more discriminative than *ROR*. Fig.1 also shows that the distinction /x/-/k/ is slightly better made in non-native than in native speech.

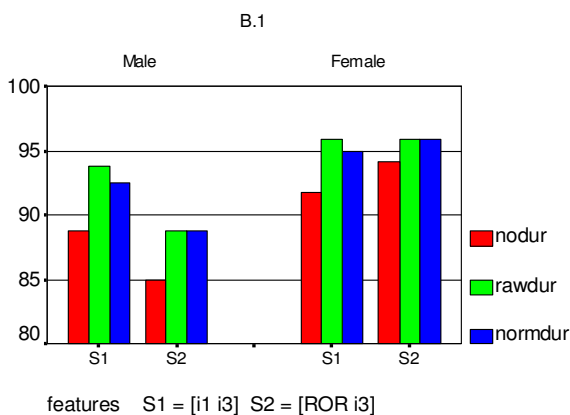


Figure 2. Correct classification %, results from exp. B.1.

Figure 2 shows how well the natively trained classifier copes with non-native speech: when the classifier is natively trained (exp. B.1) instead of non-natively (exp. A.2), the performance is almost equally high except for a small loss of approximately 2% in non-native male speech. Generally, applying non-native speech, which may be less accurately pronounced than native speech, to a natively trained model is known to be problematic. However, for this classifier this is not the case. This might be due to the fact that the relation between the steepness of the onset of the noise of fricatives and plosives is to a large degree language independent. The steepness is to a large extent responsible for the perception of the noise as a plosive, affricate or fricative. This is an example of a case where acoustic-phonetic features are more powerful than 'blind' confidence measures.

3.2 In short: classification results /A/ and /Y/

The /A/ and /Y/ LDA classifiers were trained with the three lowest formants, pitch and duration. According to the results of the A-experiments, the /A/ was correctly discriminated from /a:/ in approximately 78%-95% of all cases in the DL2N1-Nat corpus and for approximately 65%-70% in the DL2N1-NN corpus. The classification accuracy of /Y/ vs. /u,y/ was approximately 88%-100% in the DL2N1-Nat corpus and around 70% in the DL2N1-NN corpus for the A-experiments. Here, it does seem that vowels from non-native speech are less distinguishable from each other than vowels from native speech.

4 Conclusions

We can conclude that both classifiers based on an acoustic-phonetic approach and developed with a small number of relatively simple acoustic features are able to discriminate between /x/ and /k/ under different conditions with relatively high accuracy: 75%-91.7% correctness in both native and non-native speech for method I and approximately 87%-95% for method II. Furthermore, method II, i.e. the LDA classifier developed with just 2-3 features performs better than method I, i.e. the algorithm presented in [4]. Since the mispronunciation of /x/ as /k/ is a common pronunciation error made by L2-learners of Dutch, the methods presented here can be successfully employed in automatic pronunciation error detection techniques for L2-learners of Dutch.

References

- [1] Kim, Y., Franco, H. and Neumeyer, L. (1997). Automatic pronunciation scoring of specific phone segments for language instruction. *Proc. Eurospeech 1997*. Rhodes, Greece, 645-648.
- [2] Neri, A., Cucchiari, C., Strik, H. (2004). Segmental errors in Dutch as a second language: how to establish priorities for CAPT. This volume.
- [3] Cucchiari, C., Strik, H., Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America* 107,2: 989-999.
- [4] Weigelt, L.F., Sadoff, S.J. and Miller, J.D. (1990). The plosive/fricative distinction: The voiceless case. *Journal of the Acoustical Society of America* 87,6: 2729-2737.
- [5] <http://www.praat.org>