Automatic detection of frequent pronunciation errors made by L2-learners

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

[1]TNO Human Factors, Soesterberg, The Netherlands
[2]Radboud University Nijmegen, The Netherlands
[3]University of Stellenbosch, South Africa
khiet.truong@tno.nl, {A.Neri, C.Cucchiarini, H.Strik}@let.ru.nl, fdw@sun.ac.za

Abstract

In this paper, we present an acoustic-phonetic approach to automatic pronunciation error detection. Classifiers using techniques such as Linear Discriminant Analysis and Decision Trees were developed for three sounds that are frequently pronounced incorrectly by L2-learners of Dutch: /ɑ/, /ʏ/ and /x/. This paper will focus mainly on the problems with the latter phoneme. 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 second language (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 pronunciation errors in the speech of L2-learners. This paper is about developing and training classifiers for automatic detection of such pronunciation errors.

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-based confidence measures also have the disadvantage that they are not very accurate predictors of pronunciation quality at the phone level: the average human-machine correlations between the scores given by humans and the scores given by the machine 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 generally are computed on the basis of the same set of features for all the speech sounds, without focusing on the specific acoustic-phonetic characteristics of individual sounds.

Given the disadvantages of methods based on confidence measures, we have been looking for approaches that allow combining automatic speech recognition with sound-specific features, and that may 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 hopefully, to achieve higher error detection accuracy and human-machine agreement. More specificity is achieved in two ways. First, by examining the acoustic differences between the correct sound and the corresponding incorrect realization of that sound 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) and Decision Trees.

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 /ɑ/, /ʏ/ and /x/ are often mispronounced by non-native speakers, and similar erroneous realizations are generally observed across speakers with different mother tongue. Next, acoustic differences between correct and incorrect realizations of sounds were examined, which resulted in the selection of a number of promising features (section 2.2). 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. Classifiers were developed for each pronunciation error of /ɑ/, /ʏ/ and /x/. In this paper, we will focus on the /x/-classifier. A short summary of the results for /ɑ/ and /ʏ/ 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 10 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 an automatic speech recognizer (HTK) using the Viterbi algorithm.

The native part of the corpus (referred to as DL2N1-NS) consists of speech from 4 speakers of Standard Dutch and 16

speakers of regional varieties of Dutch. The non-native part (referred to as DL2N1-NNS) consists of speech from 60 non-native speakers. This non-native group is varied with respect to mother tongue and proficiency level in Dutch.

For the classification experiments, all material was divided into training (75%) and test data (25%). Furthermore, the material was divided into male 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 a part of DL2N1-NNS (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 (the agreement between annotators was generally high, see [2]). On the basis of this survey we decided to select the segmental pronunciation errors shown in Table 1 for the present study.

	Target Phon. + example in Dutch	Mispronounced as
Most frequent for vowels	/ɑ/ 'kat'	/a/
	/ʏ/ 'put'	/u/ or /y/
Most frequent for consonants	/x/ 'goed'	/k/ or /g/

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

Since in the annotated material of the non-native speakers 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 /ɑ/, /ɵ/ 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 the numbers of tokens used for the /x/-/k/ classifier, see Table 2). Note that we have more training material for female than for male speakers (see Table 2).

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

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

We did not include the /g/, since we did not have enough training material for this uncommon sound in Dutch. Separate classifiers for the 3 errors were trained, i.e. one to discriminate /ɑ/ from /a/, one for /ɵ/ versus /u, y/, and one for /x/ versus /k/. By training these pair-wise classifiers using only correct realizations, the detection is focused on the most obvious mispronunciations.

2.2. Method

This section gives an overview of the /x/-/k/ classifiers as well as the experiments that were conducted to evaluate their classification performance.

2.2.1. Method I – 'Weigelt algorithm'

In [4], a deterministic 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 24 ms window, *n*, is shifted over the acoustic signal at 1 ms intervals and for each window *n* the amplitude is measured by computing the logarithm of the Root-Mean-Square over window *n*:

$$E_n = 20*log10(RMS_n/0.00002) \quad (1)$$

ROR is then computed:

$$ROR_n = (E_n - E_{n-1}) / \Delta t \quad (2)$$

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 below 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, in [4] four criteria were used to distinguish non-significant ROR peaks from significant ROR peaks. Since the fourth criterion appeared too strict in our case, we only used three of these criteria: 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 49ms 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 fail, the peak is not significant and the consonant is classified as a fricative. If the peak is significant and its ROR value is above a predetermined ROR threshold, 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 - LDA

In addition to the algorithm in [4], another method was developed that uses Linear Discriminant Analysis (LDA) to discriminate /x/ from /k/. For this method, a number of potentially discriminative acoustic features were employed in this LDA method: duration, highest ROR value and amplitude. Duration, either raw or normalized, was chosen because fricatives are usually longer than plosives. 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 by multiplying the articulation rate per speaker by segment duration:

$$\text{normalized duration} = \text{art.rate} * \text{segment duration} \quad (3)$$

The highest ROR peak was taken, irrespective of its significance. Additionally, 4 amplitude measurements were taken to model the amplitude contour: *i1* at 5ms before the highest ROR peak, and *i2*, *i3*, *i4* at 5, 10, 20ms, respectively, after the highest ROR peak. In total, six features were examined in the LDA method: *ROR*, *i1*, *i2*, *i3*, *i4,* in combination with duration (either raw '*rawdur*', normalized '*normdur*', or not used at all '*nodur*').

All acoustic measurements (automatically obtained using Praat [5]) were based on the automatic segmentation obtained with HTK.

2.2.3. Experiments A and B

Two types of experiments were carried out for each of the two methods. In experiment A, there is no mismatch between training and test data: in exp. A.1 we train and test on native data and in exp. A.2 we train and test on non-native data (see Table 3). Experiments A.1 and A.2 were mainly carried out to test whether the methods developed were able to discriminate between /x/ and /k/ in native and non-native speech, and to examine the relative importance of each feature in the LDA method.

In experiment B, there is a mismatch between training and test data: we train classifiers with native speech and test them on non-native speech. The goal of this experiment was to investigate how a classifier trained with native (mother tongue) data would cope with non-native speech, i.e. how does the performance of a classifier trained on native data (exp. B.1) compare to the performance of a classifier trained on non-native data which is already adapted to non-native speech (exp. A.2)?

Exp.	Training	Test
A.1	DL2N1-NS	DL2N1-NS
A.2	DL2N1-NNS	DL2N1-NNS
B.1	DL2N1-NS	DL2N1-NNS

Table 3: Experiments with different training and test conditions.

3. Results

3.1. Classification results /x/

3.1.1. Method I – 'Weigelt algorithm'

We first trained the algorithm introduced in [4] on native and 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 consonants considered here.

In Table 4, the classification results obtained with this method under different training and test 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/ were correctly classified.

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

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

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

3.1.2. Method II - LDA

The second method uses LDA as classification technique to discriminate /x/ from /k/. Experiments with LDA-analyses made it clear that not all six features (*ROR, i1, i2, i3, i4,* and *nodur/rawdur/normdur*) were needed to discriminate between /x/ and /k/ (LDA offers a number of ways of pruning away less significant features). Only 2 or 3 features ([i1 i3] or [ROR i3] with duration optionally added) were sufficient to obtain classification results ranging from approximately 85% to 95% (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 do not produce a length difference between /x/ and /k/, whereas native speakers of DL2N1-NS usually do.

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], implying that (in combination with *i3*) *i1* is more discriminative than *ROR*. Next to *ROR*, there are more alternatives to model the gradual or abrupt rise of amplitude that have not been explored in this study, e.g. the absolute difference between *i1* and *i3*. Fig. 1 also shows that the distinction /x/-/k/ is slightly better made in non-native than in native speech.

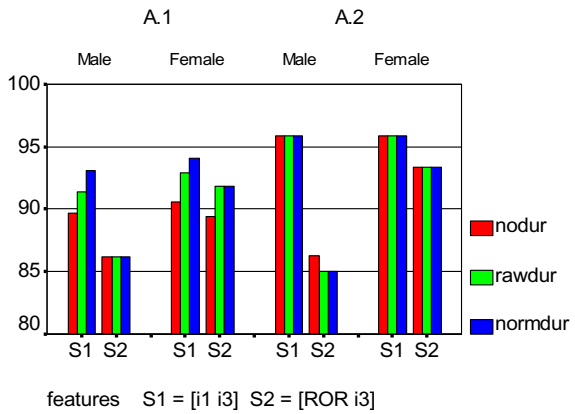

Figure 1: Correct classification in %, left = exp. A.1, right = exp. A.2.

Furthermore, the accuracy is slightly better or equal when raw duration is used instead of normalized duration. The effect of normalized duration (for articulation rate) can be absent or small because the differences between the articulation rates of the two groups of native and non-native speakers used in this study may not have been as large as expected.

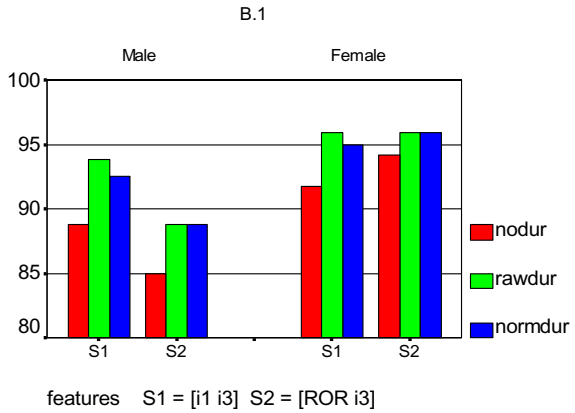

Figure 2: Correct classification in %, results from exp. B.1.

Generally, classifying non-native speech, which may be less accurately pronounced than native speech, with native models 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 plosive, affricate or fricative. This is an example of a case where an approach based on error-specific acoustic-phonetic features reaches accuracy levels that are higher than those reported in most studies using confidence measures for which it is more difficult to use specific details of the errors.

3.2. In short: classification results /ɑ/ and /ʏ/

The /ɑ/ and /ʏ/ LDA classifiers were trained with the three lowest formants, pitch and duration. According to the results of the A-experiments, the /ɑ/ was correctly discriminated from /a/ in approximately 78%-95% of all cases in the DL2N1-NS corpus (exp. A.1) and for approximately 65%-70% in the DL2N1-NNS corpus (exp. A.2). The results of exp. B.1 for /ɑ/ vs /a/ range from approximately 63% to 68%. The classification accuracy of /ʏ/ vs /u, y/ was approximately 88%-100% in the DL2N1-NS corpus (exp. A.1) and around 70% in the DL2N1-NNS corpus (exp. A.2) for the A-experiments. Exp. B.1 shows results ranging from 70% to 75% for /ʏ/ vs /u, y/. Here, it does seem that vowels from non-native speech are less distinguishable from each other than vowels from native speech (compare exp. A.1 to A.2 and B.1). For further details see [6].

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% correct in both native and non-native speech using the adjusted 'Weigelt algorithm' and approximately 87%-95% for LDA classification. Furthermore, the LDA classifier developed with just 2-3 features performs better than the 'Weigelt 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.

5. Acknowledgements

This research was carried out at the Radboud University Nijmegen within the framework of the MA thesis of the first author [6], who studied at Utrecht University. We would like to thank our colleagues at Radboud University Nijmegen for their comments on this paper.

6. References

[1] Kim, Y., Franco, H. and Neumeyer, L. "Automatic pronunciation scoring of specific phone segments for language instruction", Proc. Eurospeech, Rhodes, Greece, 645-648, 1997.

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

[3] Cucchiarini, C., Strik, H. and Boves, L. "Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology", J. Acoust. Soc. Amer., Vol. 107, 989-999, 2000.

[4] Weigelt, L.F., Sadoff, S.J. and Miller, J.D. "The plosive/fricative distinction: The voiceless case", J. Acoust. Soc. Amer., Vol. 87, 2729-2737, 1990.

[5] Boersma, P. and Weenink, D. "Praat: doing phonetics by computer (Version 4.3.01) [Computer program]", http://www.praat.org, 2005.

[6] Truong, K. "Automatic pronunciation error detection in Dutch as second language: an acoustic-phonetic approach", Master Thesis, Utrecht University, 2004. http://asterix.library.uu.nl/files/scrol/r30/