AUTOMATIC EVALUATION OF DUTCH PRONUNCIATION BY USING SPEECH RECOGNITION TECHNOLOGY

Catia Cucchiarini   Helmer Strik   Lou Boves

Dept. of Language & Speech
Nijmegen University The Netherlands

Abstract -

The ultimate aim of the research reported on in this paper is to develop a system for automatic assessment of foreign speakers' pronunciation of Dutch. The aim of the experiment described here was to determine whether pronunciation ratings assigned by human experts could be predicted on the basis of scores calculated by an automatic speech recognizer. To this end 20 native and 60 non-native speakers of Dutch read ten phonetically rich sentences over the telephone. The automatic speech recognizer was trained with read speech of 4019 Dutch subjects with varying regional accents. The results show that the human scores can be accurately predicted, even in the case of telephone speech. Analysis of the various types of human ratings and automatic measures provides more insight into the relationship between human and machine scores and indicates how the automatic measures can be further improved to achieve even greater predictive power.

1 Introduction

Developing computer tests for productive language skills such as speaking and writing is difficult because of the open-ended nature of the input. Recent advances in speech recognition research seem to suggest that there are possibilities of using computers to test at least some aspects of oral proficiency. [2, 7, 3, 8] describe automatic methods for evaluating English pronunciation. In this paper we report on an experiment that was aimed at determining whether scores obtained by means of an automatic speech recognizer correlate with human pronunciation scores of spoken Dutch. In doing so, we have analyzed both the automatic and the human experts' scores in detail.
In the methods for automatic pronunciation assessment developed so far [2, 8] different machine measures have been used: HMM log-likelihood scores, timing scores, phone classification error scores and segment duration scores. Recently, also phone log-posterior probability scores have been investigated by

[6]. In all these studies the validity of machine scores is established by comparing them with pronunciation scores assigned by human experts who are asked to assign a global pronunciation score to each of the several sentences uttered by each speaker. The scores for all the sentences by one speaker are then averaged to obtain an overall speaker score. Alternatively, the total set of sentences can be scored as a single item. Of the four measures used in [8], segment duration scores show the highest degree of correlation with human-assigned pronunciation scores (0.86). However, [6] found that phone log-posterior probability scores are even better predictors of human scores. Attempts to improve the correlations at the sentence level by combining different machine scores led to an additional 7% increase in correlation [6]. The trend in this kind of research is to look for machine measures that best correlate with human scores. In this attempt little is done to try and understand the nature of the correlation between machine scores and human scores, while this would certainly be very useful for improving automatic pronunciation assessment. Non-native speech can deviate from native speech in various aspects such as fluency, syllable structure, word stress, intonation and segmental quality. In the literature, considerable attention has been paid to the relative importance of the various aspects of speech quality for pronunciation assessment [1, 4]. The fact that human scores depend on several speech characteristics may be problematic when such scores are used as benchmark for automatic measures of speech quality. For this reason, in the present study more specific pronunciation ratings were collected along with global ratings of pronunciation quality. We asked the human raters to explicitly assess segmental quality, fluency and speech rate, in addition to overall pronunciation quality.

The present experiment includes ratings of native speech of two kinds: standard speech and speech with different regional accents. The presence of native-produced sentences might facilitate judgments of non-native speech [5]; and it is interesting to know how native regional accents are evaluated relative to the speech of foreigners.

Throughout the experiment telephone speech is used, since in the near future automatic tests to be administered over the telephone will be required for different applications. In [2] telephone quality was simulated by using 200-3600 Hz band-limited speech, but this is different from real telephone speech.

2 Aims of the present study

Given the successful attempts at developing automatic pronunciation testing systems for English, we decided to develop a similar test for assessing foreign speakers' pronunciation of Dutch. To this end we used the automatic speech recognizer developed at the University of Nijmegen. Some of the information concerning this recognizer is provided below and in [10]. The first aim of our experiment is to determine to what extent scores computed by our speech recognizer can predict pronunciation scores assigned by human experts. Furthermore, we wanted to determine whether asking the human experts to assign specific ratings of pronunciation quality along with global ratings would enhance our understanding of the relation between human scores and machine scores. The last aim of this experiment was to determine how real telephone speech would fare in an experiment of this kind.

3 Method

3.1 Speakers

The speakers in this experiment are 60 non-native speakers (NNS), 16 native speakers (NS) and 4 speakers of the standard language (SDS). The NNS were selected on the basis of language background (9 language groups), proficiency (3 levels) and sex. The NS were selected according to region of origin (4 regions) and sex. The four speakers of Standard Dutch (two males and two females) were selected on the basis of scores obtained in previous experiments in which the degree of standardness had been evaluated.

3.2 Speech material

Each speaker read two sets of five phonetically rich sentences. In preparing the sentences, the following criteria were adopted:

- the sentences should be meaningful, not sound strange and not contain foreign words or names, nor unusual words which NNS are unlikely to be familiar with;
- the content of the sentences should be as neutral as possible. They should not contain statements concerning characteristics of particular countries or nationalities;
- each set of five sentences should contain all phonemes of Dutch at least once.

The average duration of each set is 30 s. With two sets this amounts to one minute of speech per speaker. The sentences were read over the telephone. As the recording system was connected to a Euro-ISDN line, the input signals consist of 8 kHz 8 bit A-law coded samples. The subjects called from their homes or from telephone booths, so that the recording conditions were far from ideal. All speech material was checked and orthographically transcribed before being used for the experiment.

3.3 Raters

The raters involved in this experiment are three expert phoneticians with considerable experience in judging pronunciation and other speech and speaker characteristics. A high level of expertise was required because the raters had to evaluate specific aspects of pronunciation quality. The rating experiment comprised two sessions held on different days. In session 1 the raters assigned overall pronunciation scores, while in session 2 the specific scores were given. Scores were not given to individual sentences but to sets of five phonetically rich sentences. The 80 speakers were proportionally assigned to the three raters. Each rater judged 20 NNS, 6 NS (2 NS were evaluated twice) and all 4 SDS. Overall pronunciation quality, segmental quality and fluency were rated on a scale ranging from 1 to 10. A scale ranging from -5 to +5 was used to assess speech rate. Per session each rater scored 52 unique sets plus 44 sets that were added to calculate intra-rater and inter-rater reliability. Each time the order of the sets was randomized.

3.4 Automatic measures

The speech recognizer described in [10] was used. It was trained with 38 context-independent phone models, using continuous mixture density HMMs. The recognizer was trained with 18,000 phonetically rich sentences from 4019 speakers of the Polyphone database [9]. From the recognizer output the following measures were calculated:

tdur1 = total duration of speech (no pauses)
tdur2 = total duration of speech plus pauses
MSD = mean segment duration (tdur1/N-segments)
ROS = rate of speech (N-segments/tdur2)
LL = global log-likelihood (sum of LLs for individual words)

4 Results

4.1 Human scoring

Both intra-rater and inter-rater reliability coefficients α were very high ($>$ .95), except for the speech rate scores, where two raters had intra-rater reliabilities in the order of .75. Since natives consistently received higher scores, their presence could have inflated the reliability scores. However, reliabilities remained high ($>$ .91) when they were computed within the group of non-natives.

Table 1 shows the correlations between the scores on the four expert scales. It is evident that Segmental quality is almost identical to Overall pronunciation quality, but that the temporal measures are good predictors of Overall and Segmental quality too. This is in accordance with informal observations of many teachers, who report that pupils who have a low proficiency level combine disfluencies and mispronunciations.

	Overall	Segmental	Fluency	Speech
		quality		rate
Overall	1.00	0.99	0.85	0.70
Segmental quality		1.00	0.83	0.69
Fluency			1.00	0.82
Speech rate				1.00

Table 1: Correlations between the different scales

4.2 Automatic scoring

The correlations between the various automatic measures are shown in Table 2. Obviously, all correlations are very high, so that we must conclude that all measures address essentially the same characteristics of the speech.

	tdur1	tdur2	MDS	ROS	LL
tdur1	1.00	0.95	0.98	-0.96	0.94
tdur2		1.00	0.91	-0.96	0.98
MSD			1.00	-0.95	0.89
ROS				1.00	-0.94
LL					1.00

Table 2: Correlations between the various automatic scores

4.3 Automatic scoring and human scoring

Correlation coefficients were calculated between the four types of human scores and the five automatic measures. The results (corrected for attenuation) are presented in Table 3. As appears from Table 3, all correlations between automatic and human scores are high. The automatic measure that shows the highest correlations with the human scores is LL. Among the human-assigned scores, Fluency shows the highest correlations with the automatic scores.

The fact that aspects of pronunciation quality regarding speech timing, such as Fluency and Speech rate, are more highly correlated with automatic scores related to utterance duration than the scores on Overall pronunciation and Segmental quality reveals that the raters did their job properly. When asked to rate fluency and speech rate, they indeed paid attention to these aspects of speech timing. In other words, the high correlations between the four types of human-assigned scores (see Table 1) are most probably due to the fact that these aspects of pronunciation quality are indeed correlated with each other.

	Overall	Segmental quality	Fluency	Speech rate
tdur1	-0.74	-0.70	-0.90	-0.82
tdur2	-0.73	-0.68	-0.90	-0.82
MSD	-0.71	-0.67	-0.88	-0.81
ROS	0.76	0.72	0.92	0.83
LL	-0.79	-0.73	-0.91	-0.79

Table 3: Correlations between the automatic measures and the human scores

5 Discussion and conclusions

In this paper we have reported on an experiment aimed at determining whether pronunciation scores assigned by human experts can be predicted on the basis of scores produced by an automatic speech recognizer. The analyses of the human scores revealed that high levels of reliability were achieved, intrarater as well as interrater, in different conditions and for different scales. Since the human ratings appeared to be reliable, they can safely be used as a reference for the automatic scores.

The results show that overall pronunciation scores can be predicted with a considerable degree of accuracy on the basis of automatic measures. All correlations between Overall pronunciation and the automatic scores are high; the highest correlation (0.79) is found for LL. This might seem rather surprising, since in previous research [8] log-likelihood turned out to be no good predictor of overall pronunciation. However, it should be pointed out that in this experiment all automatic scores, even LL, turned out to be highly correlated with each other. The fact that there is a high correlation between LL and Overall pronunciation can thus be misleading. Inspection of the correlation between LL and Overall pronunciation revealed that the association is mostly due to the close relation between LL and utterance duration. This strong dependence of LL on utterance duration is probably due to the way in which LL is calculated by our system at the moment: the LL for the whole utterance is calculated by summing the LLs of the individual words. This strong dependence is reflected most clearly in the extremely high correlation of LL with tdur1 (of 0.94), the latter being the total duration of the utterances (i.e. of all words without the pauses). This confirms the suggestion that some kind of normalization, e.g. by computing likelihood ratios, is essential to approximate the intuitive concept of 'segmental quality' or 'overall pronunciation quality' with scores obtained from an automatic speech recognizer.

It is of interest to study the relations between automatic scores and human scores in more detail, by analyzing the 'factorial' composition of the latter. By using the specific pronunciation scores it became clear that Overall pronunciation is most influenced by Segmental quality, which is the human measure that can be predicted most poorly on the basis of the machine scores. Even

log-likelihood (LL), which was intended to be the automatic measure most closely related to Segmental quality, is highly correlated with utterance duration.
Another aspect in which our study differs from previous ones is that telephone speech was used. People were simply asked to dial a certain number, and they were free to select time, place and location. Consequently, the resulting acoustic registrations differ in many ways from those made in a studio or a (usually quiet) office environment. Here we will mention only the most relevant ones.
First of all, in telephone speech only the bandwith of 300 - 3400 Hz is used. Second, not just one high quality microphone was used, but many different telephone microphones. Finally, and probably most important, relatively high level acoustic background signals are frequently present, which is usually not the case with laboratory speech. We do consider these conditions as 'normal and realistic', in the sense that later on, when this technology will be used in applications over the telephone, conditions will most probably be similar. However, it should be underlined that these conditions make automatic speech recognition more difficult.

To conclude, the results of this experiment are very promising since they show that pronunciation scores assigned by human experts can be accurately predicted on the basis of measures computed by a speech recognizer. Furthermore, these results indicate how the machine scores could be improved so as to obtain an even greater predictive power. Finally, the fact that these results were obtained with telephone speech under 'normal and realistic' conditions, makes them even more promising.

Acknowledgements

This research was supported by SENTER (which is an agency of the Dutch Ministry of Economic Affairs) under the Information Technology Programme, the Dutch National Institute for Educational Measurement (CITO), Swets Test Services of Swets & Zeitlinger and PTT Telecom. The research of Dr. H. Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences.

References

[1] Anderson-Hsieh, J., R. Johnson and K. Koehler "The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure," *Language Learning*, Vol. 42, pp. 529-555, 1992.

[2] J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub "Automatic evaluation and training in English pronunciation," in *Proc. Int.*

Congress on Spoken Language Processing (ICSLP) '90, 1990, pp. 1185-1188.

[3] M. Eskenazi "Detection of foreign speakers' pronunciation errors for second language training - preliminary results," in *Proc. Proc. Int. Congress on Spoken Language Processing (ICSLP) '96*, 1996, pp. 1465-1468.

[4] M.J. Munro "Nonsegmental factors in foreign accent," *Studies in Second Language Acquisition*, Vol. 17, pp. 17-34, 1995.

[5] J. Flege and K. Fletcher "Talker and listener effects of perceived foreign accent," *Journal of the Acoustical Society of America*, Vol. 91, pp. 370-389, 1992.

[6] H. Franco, L. Neumeyer, Y. Kim and O. Ronen "Automatic pronunciation scoring for language instruction," in *Proc. Int. Congress on Acoustics, Speech and Signal Processing (ICASSP) 1997*, pp. 1471-1474.

[7] Hiller, S., E. Rooney, R. Vaughan, M. Eckert, J. Laver, and M. Jack "An automated system for computer-aided pronunciation learning," *Computer Assisted Language Learning*, Vol. 7, pp. 51-63, 1994.

[8] L. Neumeyer, H. Franco, M. Weintraub, and P. Price "Automatic text-independent pronunciation scoring of foreign language student speech," in *Proc. Proc. Int. Congress on Spoken Language Processing (ICSLP) '96*, Philadelphia, pp. 1457–1460.

[9] den Os, E.A., T.I. Boogaart, L. Boves and E. Klabbers "The Dutch Polyphone corpus," in *Proc. ESCA 4th European Conference on Speech Communication and Technology: EUROSPEECH 95*, Madrid, pp. 825-828.

[10] H. Strik, A. Russel, H. van den Heuvel, C. Cucchiarini and L. Boves "A spoken dialogue system for the Dutch public transport information service" *International Journal of Speech Technology*, Vol. 2, pp. 119-129, 1997.