AUTOMATIC ASSESSMENT OF FOREIGN SPEAKERS’ PRONUNCIATION OF DUTCH

C. Cucchiarini and L. Boves

University of Nijmegen, Dept. of Language & Speech
P.O. Box 9103, 6500 HD Nijmegen, The Netherlands
tel: 31-24-3615785, fax: 31-24-3615939
{Cucchiarini, Boves}@let.kun.nl, http://lands.let.kun.nl/staff/

ABSTRACT

The aim of the research reported on here is to develop a system for automatic assessment of foreign speakers’ pronunciation of Dutch. In this paper similar studies carried out for English are first examined. Subsequently, suggestions are made for partly improving the methodology that is usually adopted in research on automatic pronunciation assessment. Finally, an experiment is presented in which automatic scores of telephone speech produced by native and nonnative speakers are compared with scores assigned by human raters. The approach used in this experiment is compared with those of previous studies.

1. INTRODUCTION

Every year in the Netherlands lots of foreigners take part in examinations aimed at testing their proficiency in Dutch. In order to achieve greater efficiency and lower costs, attempts are being made to automate at least part of the testing procedure. Automatic testing of receptive skills such as reading and listening appears to be relatively simple, because the response tasks that are often used -multiple choice, matching and cloze- are easy to score. Developing computer tests for productive skills such as speaking and writing is more difficult because of the open-ended nature of the input. On the other hand, it is precisely for testing these latter skills that extremely high costs are incurred, because the task human raters have to carry out is very time-consuming.

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. For instance, [1, 2, 3, 4] describe automatic methods for evaluating English pronunciation. In the wake of the success obtained in developing pronunciation tests for English based on speech recognition technology, we started a research project which aims at developing a similar system for automatic assessment of foreign speakers’ pronunciation of Dutch. In this project the University of Nijmegen collaborates with the Dutch National Institute for Educational Measurement (CITO).

In this paper we first consider some of the various systems for automatic pronunciation scoring that have been developed so far. Subsequently, we pay attention to the type of human pronunciation scoring that is usually adopted in studies of this kind. We then consider the importance of human scores in this research, because they are used as benchmark for validating the machine scores. In the final part of the paper we describe the approach adopted in our study and discuss how this research relates to similar studies carried out elsewhere.

2. PREVIOUS STUDIES

2.1. Automatic scoring of pronunciation quality

In the various methods for automatic pronunciation assessment developed so far (e.g. [1, 4]) different machine measures have been used for automatic scoring: 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 [5].

In all these studies, the validity of machine scores is established by comparing them with pronunciation scores assigned by human experts (human scores). In general, the raters are asked to assign a global pronunciation score to each of the several sentences uttered by each speaker (sentence level rating). The scores for all the sentences by one speaker are then averaged so as to obtain an overall speaker score (speaker level rating) (see[4, 5]). Although this procedure may seem logical at first sight, there are some problems with it.

The scores assigned by one and the same rater to different sentences uttered by one and the same speaker may differ as a function of segmental makeup. For example, if a stigmatizing sound (shibboleth) is present in one sentence, the score for that sentence may be considerably lower than that of other sentences that do not contain that specific sound. It may even be the case that were the rater to assign a pronunciation score for the speaker instead of for the sentence, (s)he would be heavily influenced by the presence of that stigmatizing sound such as to assign a very low overall speaker score [6]. If this were the case, then the average score computed over all sentences by one speaker would not take account of the effect of the shibboleth sound. This seems to suggest that if the researcher is interested in pronunciation scores at the speaker level, (s)he should have the human raters listen to a balanced set of sentences by each speaker and then assign an overall pronunciation score to each speaker. The reason for this is that arithmetically derived speaker scores, obtained by averaging the relative sentence scores, may not reflect the raters’ speaker judgements.

In the studies mentioned above, correlations between automatic scores and human scores appear to be higher at the speaker level than at the sentence level. Sentence-level correlations are all very low, whereas at the speaker level considerable differences are observed between the various measures (HMM log-likelihood scores, timing scores, phone classification error scores and segment duration scores). Of the four measures used in [4], segment duration scores show the highest degree of correlation with human-assigned pronunciation scores (0.86). However, [5] found that phone log-posterior probability scores are even better predictors of human pronunciation scores (the correlation between phone log-posterior probability and human scores turns out to be 0.88). Attempts to improve the correlations at the sentence level by combining different machine scores led to an additional 7% increase in correlation [5].

Quite clearly the trend in this kind of research is to look for machine measures that best correlate with human scores. What is striking is that 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. For example, there seems to be a mismatch between the knowledge available on machine scores and that concerning human scores. While the machine scores are relatively clear, that is to say that it is known what each measure stands for, very little is known about the human scores. In the above-mentioned studies, the expert raters were asked to give a global rating of pronunciation quality. However, research on pronunciation evaluation has revealed that scores of pronunciation quality may be affected by a great variety of speech characteristics, as will be explained in the following section.

2.2. Human scores of pronunciation quality

Nonnative speech can deviate from native speech in various aspects such as fluency, syllable structure, word stress, intonation and segmental quality. When native speakers are asked to score nonnative speech on pronunciation quality, their scores are usually affected by more than one of these areas. In the literature, considerable attention has been paid to the relative importance of the various aspects of pronunciation quality for intelligibility [7, 8, 9, 10, 11, 12 and 13]. Research aimed at investigating the relationship between native speaker ratings of nonnative pronunciation and deviance in the various areas of speech quality has revealed that each area affects the overall score to a different extent [13].

These findings suggest that global ratings of pronunciation quality assigned by human raters have a complex structure. This may be problematic when such scores are used as benchmark for automatically produced measures of speech quality, because one simply does not know what the human scores stand for. It is our impression that questions such as “What do raters exactly evaluate?” and “What influences their judgements most?” should be taken into consideration when trying to develop machine measures that best approach human pronunciation scores.

2.3. Human pronunciation scores as benchmark for automatic scores

From the previous section it appears that human experts, when rating pronunciation quality, may pay attention to all sorts of different speech characteristics such as fluency, word stress, intonation, segmental quality or even other aspects of speech we might not have thought of. In order to understand how machine scores are correlated with human scores, it might be very useful to know what expert raters exactly pay attention to. Furthermore, it should be noted that the scores produced by a speech recognizer, such as HMM log-likelihood scores, phone log-posterior probability scores, timing scores and phone classification error scores (see also [4 and 5]) do not cover all the above-mentioned areas. Therefore, in order to obtain a more clear-cut idea of how automatic scores agree with human ratings, it would be better to ask the human raters to judge those aspects of pronunciation quality of which we know that they can be evaluated by both man and machine. Moreover, studying different aspects separately would certainly contribute to our understanding of this complex relationship.

3. 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. Further details can be found in [15].

For the reasons mentioned above, in our study of automatic pronunciation assessment we opted for an approach that differs from those adopted in previous studies in various respects. The differences between our method and those of other studies are discussed below.

3.1. How this study differs from previous ones

As mentioned above, the use of global ratings of pronunciation quality in research on automatic pronunciation assessment is questionable. For this reason, we decided to study one aspect of pronunciation at a time, instead of collecting global ratings of pronunciation quality (for a similar approach, see [2]). Of all the aspects in which nonnative speech can deviate from native speech, segmental quality and prosody have received the greatest attention [13]. Therefore we decided to start with one of these two aspects. Segmental quality is the first area we have selected for investigation. The experiment concerning the assessment of segmental quality will be described below. Other aspects such as word stress, intonation and fluency will be addressed in following experiments.

Another feature that distinguishes the experiment reported on here is that the human raters are not asked to assign separate sentence scores that will be averaged to obtain an overall speaker score. Instead, the raters will judge the pronunciation of each speaker on the basis of two sets of phonetically rich sentences.

Furthermore, this experiment is characterized by the fact that it is not limited to assessing nonnative speech, but it also concerns native speech of two kinds: standard speech and speech with different regional accents. The first reason for doing this is that the presence of native-produced sentences facilitates judgements of nonnative speech [16]. Second, it is interesting to know how native strong regional accents are evaluated in the same experiment, and whether human raters score them in the same way as the machine does.

Finally, another characteristic of this experiment is that telephone speech is used. The rationale behind this is that in the near future automatic tests to be administered over the telephone will be required for different applications. In one study that we know of [1] telephone quality was simulated by using 200-3600 Hz band-limited speech. Of course this is not the same thing as using real telephone speech.

3.2. Experimental setup

In this experiment the speech recognizer described in [15] was used. The system was connected to an ISDN line. Therefore, the input signals consist of 8 kHz 8 bit A-law coded samples. Feature extraction is done every 10 ms for frames with a width of 16 ms. The first step in feature analysis is an FFT analysis to calculate the spectrum. Next, the energy in 14 mel-scaled filter bands between 350 and 3400 Hz is calculated. Apart from these 14 filterbank coefficients the 14 delta coefficients, log energy, and slope and curvature of the energy are also used. This makes a total of 31 feature coefficients.

The CSR uses acoustic models (HMMs), language models (unigram and bigram), and a lexicon. The lexicon contains orthographic and phonemic transcriptions of the words to be recognized. The continuous density HMMs consist of three segments of two identical states, one of which can be skipped.

The CSR was trained by using part of the Polyphone database (see [14]). This corpus is recorded over the telephone and consists of read and (semi-)spontaneous speech of 5000 subjects with varying regional accents. For each speaker 50 items are available. Five of these 50 items are the so-called phonetically rich sentences, which contain all phonemes of Dutch at least once, while the more frequent phonemes occur more often. This part of the database was used for training.

For the present experiment two sets of five phonetically rich sentences were prepared. In each set all phonemes of Dutch appear at least once. These sentences were read over the telephone by 68 speakers. Data collection proceeded in the same way as was done for the Polyphone database. The sentences were then processed by the speech recognizer. HMM log-likelihood scores were calculated for all sentences. Speaker scores were obtained by averaging the scores for the five sentences. In this case this is legitimate, because the machine is not likely to be affected by shibboleth phenomena. In computing the automatic scores, a text-dependent approach (see [4]) was adopted. This implies that knowledge about the sentences was used, for instance by applying forced alignment. On the basis of the HMM-log likelihood scores a rank ordering of the utterances was established.

Subsequently, the utterances will be evaluated by expert human raters who will be explicitly asked to score segmental quality alone, for each separate balanced set of five sentences. On the basis of the human judgments a rank ordering will again be determined. Finally, the scores calculated by the automatic speech recognizer will be compared with the scores assigned by the human experts.

3.3. Speakers

The speakers involved in this experiment are 48 nonnative speakers (NNS), 16 native speakers (NS) and 4 speakers of the standard language (SDS). The NNS were selected on the basis of the following variables:1. sex (two levels), 2. level of proficiency (three levels) and 3. mother tongue (eight levels). On the basis of these three variables a 2 x 3 x 8 factorial design is obtained. By selecting one speakers per cel a sample of 48 speakers is obtained.

The group of NS was selected according to the following variables: 1. sex and 2. region of origin (four different regions were selected as to obtain four different dialect backgrounds). The 16 NS all have a low educational level (this was done to be sure that the speakers had a regional accent).

Four speakers of Standard Dutch (two males and two females) were also included. They were selected on the basis of scores obtained in previous experiments in which the degree of standardness had been evaluated.

3.4. Speech material

Each speaker read two sets of five phonetically rich sentences. An obvious choice seemed to be to use two of the many sets of five phonetically rich sentences that had been prepared for the Polyphone database. However, many of these sentences appear to be rather difficult for learners of Dutch, for various reasons. Therefore, we decided to use some of the existing material, when possible, while in other cases new sentences were prepared. The criteria adopted in selecting the sets of sentences are the following:

- the sentences should be meaningful and should not sound strange
- the sentences should not contain unusual words which NNS are unlikely to be familiar with
- the content of the sentences should be as neutral as possible. For instance, the sentences should, preferably, not contain statements concerning characteristics of particular countries or nationalities
- the sentences should not contain foreign words or names
- the sentences should not contain long compound words which are particularly difficult to pronounce
- each set of five sentences should contain all phonemes of Dutch at least once, and, preferably, more common phonemes should apperar more than once.

The average duration of each set is 30 sec. With two sets this amounts to one minute of speech per speaker.

3.5. Rating procedure

The utterances produced by the 68 speakers will be recorded on tape and presented to a group of raters. Each rater will assign a score to each set of five sentences. This way, two scores will be obtained for each speaker. For scoring a ten-point scale will be used. In order to be able to determine intrarater reliability, part of the utterances (10%) will be judged twice by each rater. Interrater reliability will be established by comparing the scores assigned by the different raters. Half of the raters participating in the experiment will be phoneticians, while the remaining part will be speech pathologists who have experience in diagnosing learners of Dutch as a second language.

Raters will be instructed to score segmental quality alone. A list of aspects which should not be taken into consideration (word accent, intonation, speech tempo etc.) will also be provided. Segment duration should be taken into account because it concerns segmental quality.

4. CONCLUSION

In this paper we have considered some of the studies on automatic pronunciation assessment that have been published so far. We have pointed out that in most of these studies relatively little attention is paid to the human scores that are used as benchmark for machine scores. More specifically, it is our impression that researchers focus their attention on finding machine measures that correlate strongly with the human scores, whereas little is known about what these human scores exactly stand for. Therefore we have argued for the use of more detailed judgements of pronunciation quality in order to obtain better insight in the way in which human pronunciation scores correlate with machine scores of various kinds.

Furthermore, we have observed that in the studies under consideration speaker level ratings are obtained by averaging sentence level ratings. Since this procedure has little to commend itself, we have suggested that speaker level scores be collected directly from the raters. This approach has been adopted in our study on automatic pronunciation assessment.

In addition, our study is not limited to nonnative speech, but comprises accented native speech and standard speech. Finally, the fourth innovative aspect of our study is that real telephone speech is used. The results concerning the automatic scores of the recorded utterances and their comparison with human scoring will be presented at the poster session.

5. ACKNOWLEDGEMENT

This research was supported by SENTER, which is an agency of the Dutch Ministry of Economics, under the Information Technology Programme.

6. REFERENCES

[1] Bernstein, J., M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub (1990), "*Automatic evaluation and training in English pronunciation*". Proc. ICSLP '90, pp. 1185-1188.

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

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

[4] Neumeyer, L., Franco H. M. Weintraub, and P. Price (1996), "*Automatic text-independent pronunciation scoring of foreign language student speech*". Proc. ICSLP '96, pp. 1457-1460.

[5] Franco, H., L. Neumeyer, Y. Kim and O. Ronen (1997), "*Automatic pronunciation scoring for language instruction*", Proc. ICASSP 1997, pp. 1471-1474.

[6] R. van Bezooijen, personal communication

[7] James, E. (1976), "The acqqusition of prosodic features using a speech visualizer", *International Review of Applied Linguistics and Language Teaching*, Vol. 14, pp. 227-243.

[8] Johansson, S. (1978), "Studies of error gravity: Native reactions to errors produced by Swedish learners of English", Göteborg, Sweden, Acta Universitatis Gothoburgensis.

[9] van Heuven, V.J. and J.W. de Vries (1981), "Begrijpelijkheid van buitenlanders: de rol van fonische en niet-fonische factoren", *Forum der Letteren*, Vol. 22, pp. 309-320.

[10] Fayer, J. and E. Krazinsky (1987), Native and nonnative judgments of intelligibility and irritation, *Language Learning*, 37, pp. 313-326.

[11] Anderson-Hsieh, J. and K. Koehler (1988), "The effect of foreign accent and speaking rate on native speaker comprehension", *Language Learning*, Vol. 38, pp. 561-613.

[12] Boeschoten, J. (1989), "Verstaanbaarheid van klanken in het Nederlands gesproken door Turken", PhD Dissertation Leyden University

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

[14] den Os, E.A., T.I. Boogaart, L. Boves and E. Klabbers (1995), "*The Dutch Polyphone corpus*", Proc. EUROSPEECH 95, pp. 825-828, Madrid.

[15] Strik, H., A. Russel, H. van den Heuvel, C. Cucchiarini and L. Boves (1997), "A spoken dialogue system for the Dutch public transport information service", to appear in *International Journal of Speech Technology.*

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