Using speech recognition technology to assess foreign speakers' pronunciation of Dutch

Catia Cucchiarini, Helmer Strik and Lou Boves
University of Nijmegen

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, Bernstein et al. (1990), Hiller et al. (1994), Eskenazi (1996) and Neumeyer et al. (1996) describe automatic methods for evaluating English pronunciation. In 1996 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 cooperates with the Dutch National Institute for Educational Measurement (CITO), Swets Test Services of Swets & Zeitlinger and PTT Telecom.

In this paper we first describe the goals of the present experiment (section 2). We then go on to consider how this study differs from previous ones (section 3). In section 4 the methodology is described. The results of this experiment are presented in section 5. Finally, in section 6 the results are discussed and some conclusions are drawn.

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. Further details can be found in Strik et al. (1997). The first aim of the experiment reported on here 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. Another aim of this experiment was to determine whether native and nonnative speakers of Dutch are evaluated in the same way by man and machine.

3. How this study differs from previous ones

In the various methods for automatic pronunciation assessment developed so far (e.g. Bernstein et al. 1990 and Neumeyer et al. 1996) 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 Franco et al. (1997).

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 Neumeyer et al. 1996 and Franco et al. 1997). 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 make-up. 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. 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 fragments containing the whole phonetic inventory. The reason for this is that speaker scores obtained by averaging the relative sentence scores may not reflect the raters' speaker judgements. In view of this, in the present experiment the human raters were not asked to assign scores to individual sentences. Instead, the raters judged the pronunciation of each speaker on the basis of two sets of phonetically rich sentences.

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. 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 Neumeyer et al. (1996), segment duration scores show the highest degree of correlation with human-assigned pronunciation scores (0.86). However, Franco et al. (1997) 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). Sentence-level correlations, on the other hand, are all very low. Attempts to improve the correlations at the sentence level by combining different machine scores led to an additional relative increase by 7% (i.e. from 0.58 to 0.62) in correlation (Franco et al. 1997).

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 how they are calculated, very little is known about the human scores.

Research on pronunciation evaluation has revealed that scores of pronunciation quality may be affected by a great variety of speech characteristics. 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 aspects. In the literature, considerable attention has been paid to the relative importance of the various aspects of pronunciation quality for intelligibility (James 1976; Johansson 1978; van Heuven & de Vries 1981; Fayer & Krasinsky 1987; Anderson-Hsieh & Koehler 1988; Boeschoten 1989; Anderson-Hsieh, Johnson & Koehler 1992). Research aimed at investigating the relationship between native speaker ratings of nonnative pronunciation and deviance in the various aspects of speech quality has revealed that each area affects the overall score to a different extent (Anderson-Hsieh, Johnson & Koehler 1992).

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 a 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. For this reason, in the present study more specific pronunciation ratings were collected along with global ratings of pronunciation quality .

In deciding which aspects of pronunciation quality should be investigated in this experiment we took account of the fact 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 Neumeyer et al. 1996 and Franco et al. 1997) 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, we asked the human raters to judge those aspects of pronunciation quality of which we expect that they can be evaluated by both man and machine such as segmental quality, fluency and speech rate.

Furthermore, the present 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 (Flege & Fletcher 1992). 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 telephone quality was simulated by using 200-3600 Hz band-limited speech (Bernstein et al. 1990). Of course this is not the same thing as using real telephone speech.

4. Method

4.1 Speakers

The speakers involved in this experiment are 60 nonnative speakers (NNS), 16 native speakers with strong regional accents (NS) and 4 Standard Dutch speakers (SDS). The speakers in the three groups were selected according to different sets of variables, as is shown below:

1) The 60 NNS were selected on the basis of the following three variables:
   - language background (9 language groups)
   - proficiency level (3 levels)
   - sex (2)

2) The 16 NS were selected according to:
   - region of origin (4 regions)
   - sex (2)

3) The four speakers of Standard Dutch (two males and two females) were selected on the basis of the high scores they had obtained in previous experiments in which the degree of standardness of their pronunciation had been evaluated.

4.2 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 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 appear more than 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 an 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.

4.3 Human scoring

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 experiment was divided into two sessions which were held on different days. In session 1 the raters assigned overall pronunciation ratings, while in session 2 specific ratings were assigned for segmental quality, fluency and speech rate. This setup was chosen so as to ensure that the overall ratings would not be

influenced by the specific ones. 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.

The 80 speakers were proportionally assigned to the three raters. Each rater was assigned: 20 NNS, 6 NS (2 NS were evaluated twice) and all 4 SDS. Since for each speaker two sets of sentences had to be evaluated, each rater scored 40 sets of NNS, 12 of NS and 8 of SDS. Per session each rater rated 52 unique sets (40 NNS plus 12 NS) plus 44 sets that were scored by all three raters. Thus each rater assigned 96 scores for each of the 4 scales: overall, segmental quality, fluency and speech rate. Each time the order of the sets was randomized. The 44 sets that were scored by all three raters were used to calculate interrater reliability. For each rater intrarater reliability was calculated on the basis of 12 of these 44 sets that the rater in question had scored twice.

4.4 Automatic scoring

In this experiment the speech recognizer described in Strik et al. (1997) was used. Feature extraction is done every 10 ms for frames with a width of 16 ms. The first step in feature analysis is a Fast Fourier Transformation (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 continuous speech recognizer (CSR) uses acoustic models (context-independent Hidden Markov Models, CIHMMs), 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. 38 context-independent phone models were trained.

The CSR was trained by using part of the Polyphone database (den Os et al. 1995). 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. Each speaker read a different set of sentences. In this experiment speech from 4019 speakers was used for training the CSR.

As mentioned above, the human raters were asked to evaluate those aspects of pronunciation quality for which meaningful automatic correlates might be calculated. Automatic speaker scores were obtained by averaging the scores for the five sentences and for the two sets. 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 Neumeyer et al. 1996) was adopted. This implies that knowledge about the sentences was used by applying a form of forced Viterbi alignment. 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 (calculated for the whole utterance, including pauses)

5. Results

5.1 Human scoring

5.1.1 Intrarater reliability

On the basis of the sets of sentences that each rater evaluated twice (24 scores), intrarater reliability could be established. The results for the three raters are shown below:

Table 1 Intrarater reliability for 3 raters for 4 scales

	Overall	Segmental quality	Fluency	Speech rate
Rater 1	α = .97	α = .96	α = .97	α = .94
Rater 2	α = .95	α = .98	α = .94	α = .76
Rater 3	α = .99	α = .93	α = .95	α = .74

As appears from Table 1, raters 2 and 3 achieve a lower degree of reliability in scoring speech rate.

5.1.2 Interrater reliability

Interrater reliability was calculated on the basis of the 44 sets of sentences that were evaluated by all three raters. Since native speakers and in particular standard language speakers consistently receive higher scores than the nonnative speakers, their presence has the effect of increasing the correlation between the scores assigned by the three raters. For this reason, the degree of reliability was computed for three different conditions: 1. SDS NS NNS (all three groups of speakers), 2. NS NNS (without Standard Dutch speakers) and 3. NNS (only foreign speakers).

Table 2 Interrater reliability for 4 scales in 3 conditions

	Overall	Segmental quality	Fluency	Speech rate
SDS NS NNS	α = .97	α = .97	α = .96	α = .86
NS NNS	α = .96	α = .97	α = .95	α = .84
NNS	α = .89	α = .92	α = .96	α = .87

As is clear from Table 2, even in the least favourable condition (NNS), the reliability coefficients are still rather high.

5.1.3 Comparing human pronunciation scores

By comparing the overall scores with the specific ones, it is possible to establish which of the separate aspects of pronunciation quality investigated here has the greatest impact on the overall score. Because the reliability coefficient differs for the various scales, the correlation coefficients have been corrected for attenuation (Ferguson 1987: 442).
As is clear from Table 3, all correlations between the human scores are high. The highest correlation is found between the Overall scores and Segmental quality. In other words, when the raters judge overall pronunciation, they are most influenced by the quality of the segments uttered by the speaker. The fact that all correlations are high is amenable to two different interpretations: either the various aspects of pronunciation quality are indeed highly correlated with each other, in which case the raters did their job properly, or the raters failed to score the various aspects independently of each other. At this point no choice can be made between these two interpretations. A comparison of the human scores with the machine scores may throw some light on this (see section 5.3).

Table 3 Correlations between the different scales

	Overall	Segmental quality	Fluency	Speech 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

5.2 Automatic scoring

The correlations between the various automatic measures are shown in Table 4.

Table 4 Correlations between the various automatic scores

	tdur1	tdur2	MDS	ROS	LL
tdur1	1.00	0.95	0.98	-0.96	0.94
tdur2		1.00	0.91	-0.96	0.99
MSD			1.00	-0.95	0.89
ROS				1.00	-0.94
LL					1.00

All correlations are very high. In this case $\alpha = 1.00$ because in repeating the calculations exactly the same scores would be obtained. Therefore, no correction for attenuation was applied.

5.3 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 5. As appears from Table 5, all correlations between automatic and human scores are high. The automatic measure that shows the highest correlations with the human scores is ROS. Among the human-assigned scores, Fluency shows the highest correlations with the automatic scores.

These data also provide an answer to the question we posed in section 5.1.3. 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 3) are most probably due to the fact that these aspects of pronunciation quality are indeed correlated with each other.

Table 5 Correlations between the automatic measures and the human scores

	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	-0.73	-0.68	-0.89	-0.81

6. 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. Furthermore, as described in section 5.3, the results indicate that the human raters correctly evaluated the aspects of pronunciation we asked them to evaluate. Since the human ratings appeared to be reliable, they could be used as a benchmark 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, while the highest correlation (0.76) is found for ROS. All automatic scores turned out to be highly correlated with each other, the reason being that all automatic scores are related to utterance duration. The consequence is that hardly any gain in predictive power can be obtained by combining automatic scores. This was confirmed by a multiple regression analysis in which Overall pronunciation was the dependent variable. When another variable was entered in the multiple regression equation after ROS, the multiple correlation coefficient only showed a marginal increase.

If the human raters had rated Overall pronunciation alone, as was the case in many previous studies, nothing more could have been said about the correlations between the automatic measures and Overall pronunciation. However, our study differs from previous ones in that we have also collected more specific ratings of pronunciation quality. These specific ratings of pronunciation quality made it possible to gain more insight into the relations between human and automatic scores, as will be explained below.

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. In fact, when the LL scores were normalised for duration, they no longer showed any correlation with the human scores.

The fact that there is a high correlation between LL and Overall pronunciation can thus be misleading. Given the nature of the LL measure one might think that this is mainly because LL is a good measure of segmental quality. However, closer inspection revealed that this was not the case. The high correlation between LL and Overall pronunciation 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 and the pauses. This strong dependence is reflected most clearly in the extremely high correlation of LL with tdur2 (of 0.99), the latter being the total duration of the utterances (i.e. of all words plus pauses).

Subsequently, an important goal of our research will be to find a measure which is more related to segmental quality than LL. Such a measure should make it possible to predict overall pronunciation with an even higher degree of accuracy than was obtained in this experiment. This could be done by combining this measure of segmental quality with a temporal measure, like e.g. ROS.

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.

7. 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.

8. References

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.

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.

Bernstein, J., M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub (1990) *Automatic evaluation and training in English pronunciation*. Proc. Int. Congress on Spoken Language Processing (ICSLP) '90, Kobe, pp. 1185-1188.

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

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

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

Ferguson, G.A. (1987) *Statistical analysis in psychology and education*, fifth edition, McGraw-Hill book company, Singapore.

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.

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

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.

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.

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

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

Neumeyer, L., H. Franco, M. Weintraub, and P. Price (1996) *Automatic text-independent pronunciation scoring of foreign language student speech*. Proc. Proc. Int. Congress on Spoken Language Processing (ICSLP) '96, Philadelphia, pp. 1457-1460.

den Os, E.A., T.I. Boogaart, L. Boves and E. Klabbers (1995) *The Dutch Polyphone corpus*, Proc. ESCA 4th European Conference on Speech Communication and Technology: EUROSPEECH 95, Madrid, pp. 825-828.

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*.