# ASSESSMENT OF DUTCH PRONUNCIATION BY MEANS OF AUTOMATIC SPEECH RECOGNITION TECHNOLOGY

*Catia Cucchiarini, Febe de Wet, Helmer Strik and Lou Boves*

*A²RT*, Dept. of Language and Speech, University of Nijmegen, the Netherlands
{Catia,deWet,Strik,Boves}@let.kun.nl; http://lands.let.kun.nl/

## ABSTRACT

Experiments were carried out to determine whether log-likelihood ratios (LRs) can be employed to improve automatic assessment of Dutch pronunciation. Read speech of natives and non-natives was judged by three groups of expert raters and was then analyzed by means of a continuous speech recognizer. Three automatic measures were calculated, two LRs and rate of speech (ros), and then compared with the expert ratings. It appears that expert ratings of pronunciation quality can accurately be predicted on the basis of ros alone and that LRs do not contribute to better prediction. However, LRs can be useful to automatic pronunciation assessment because they can help detect fast speakers who produce totally wrong sentences.

## 1. INTRODUCTION

The eventual aim of the research reported on in this paper is to develop an automatic system of pronunciation grading for Dutch by using speech recognition technology. As in other studies [1, 2, 3], in this investigation the performance of the speech recognizer is validated against pronunciation scores assigned by human experts. Important characteristics of the present investigation as compared to previous ones are that different groups of human experts are involved as raters and that these human raters were required to evaluate several aspects of pronunciation quality.

Some of the results obtained in this study have already been reported in previous papers [4, 5]. For instance, we were able to show that normalizing the expert scores can lead to better results and greater insight and that, after normalization, the correlations between the scores of the three rater groups and the correlations between the expert scores and the automatic scores are very similar, which suggests that the results obtained are not so much dependent on the choice of the raters [4]. Furthermore, we have previously reported that automatic measures of speech quality that are related to temporal properties of speech are able to predict expert pronunciation scores with a high degree of accuracy [4, 5]: fast speakers generally receive high pronunciation ratings.

These results suggest that for our pronunciation grading system it would suffice to measure these temporal variables. Although this would be convenient, because these automatic measures can be calculated relatively easily, it is likely that these temporal measures will fail in some cases. For instance, the subject can either speak the target sentence very fast, but with a poor pronunciation, or he/she can rapidly speak another sentence than the target sentence. One might hope that an off-the-shelf continuous speech recognizer (CSR), when used cleverly, should be able to detect both problems. In [8] it was shown that likelihood ratios are a promising way to try to detect whether another utterance than the prompted one has been spoken. The ratios used in [8] are not simple to compute with an off-the-shelf CSR, if only because they require very specific anti-models to be trained. Therefore, in this paper we investigate whether other ratios, which are straightforward to compute, can contribute to automatic assessment of pronunciation quality, independent of speech rate measures. In doing so, we intend to improve our understanding of the pronunciation quality assessment process itself.

## 2. METHOD

### 2.1. Speakers and Speech Material

The speakers involved in this experiment are 60 non-native 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, such as language background, proficiency level and sex, for the NNS, and region of origin and sex for the NS. Each speaker read two sets of five phonetically rich sentences (about one minute of speech per speaker) over the telephone. The speech material was orthographically transcribed. For further details, see [5].

### 2.2. Raters

Since in this experiment specific aspects of pronunciation quality had to be evaluated (see 2.3), raters with a high level of expertise were required. Different raters seemed to qualify as pronunciation experts: phoneticians, because they are expert on pronunciation in general; teachers of Dutch as a second language (L2) for obvious reasons. However, it turned out that, in practice, pronunciation problems of people learning Dutch as L2 are usually not addressed by language teachers, but by specially trained speech therapists. Therefore, phoneticians and speech therapists were selected for this investigation. Since we could easily find a second group of speech therapists, three rater groups were eventually involved: three phoneticians (ph) with experience in judging speech and speaker characteristic and two groups of speech therapists (st1 and st2) expert on pronunciation problems of Dutch L2 learners.

### 2.3. Expert Pronunciation Ratings

The experts rated four different aspects of oral delivery in two sessions: Overall Pronunciation (OP) in session 1, and Segmental Quality (SQ), Fluency (FL) and Speech Rate (SR) on a separate occassion in session 2. We chose to have them evaluate these

aspects, because we thought these were the characteristics that could be evaluated relatively easily by both man and machine.

All raters listened to the speech material and assigned scores individually. OP, SQ and FL were rated on a scale ranging from 1 to 10. A scale ranging from -5 to +5 was used to assess SR. Since it was not possible to have all raters score all speakers (it would cost too much time and it would be too tiring for the raters) the 80 speakers were proportionally assigned to the three raters in each group. Each rater was assigned 20 NNS, 6 NS (2 NS were evaluated twice) and all 4 SDS. The scores assigned by the three raters were then combined to compute correlations with the machine scores. More detailed information concerning the rating procedure can be found in [5].

## 2.4. Automatic Assessment of Pronunciation Quality

Automatic measures were calculated by means of different versions of an HMM-based CSR (for further details about the CSR, see [6]). The training material consisted of the phonetically rich sentences of 4019 speakers from the Polyphone data base [7]. 38 context independent phone models were trained. The phonetic transcriptions used in the training were obtained by concatenating the canonical transcriptions of the words, taken from a lexicon. These phone transcriptions were also used to train phone language models (unigram and bigram). Next, transcriptions in terms of five Broad Phonetic Classes (BPCs) were obtained (vowels, liquids, nasals, fricatives and plosives) by replacing all phones by their respective BPC symbols. These BPC transcriptions were employed to train BPC models and BPC language models (again unigram and bigram). Likelihood (LH) scores were calculated with several different procedures, always with

LH1. Forced Viterbi alignment with the canonical transcriptions and the 38 monophone HMMs

as the numerator term. Different denominator terms were used, e.g.

LH2. Free phone recognition with the same 38 monophone HMMs, using the phone language models during the decoding (i.e., applying loose phonotactic constraints);

LH3. BPC recognition with HMMs for the BPC models, using the BPC language models.

The general idea is that LH1 should be positively correlated with pronunciation quality: the better the actual speech fits the canonical sequence of phone models, the better the perceived quality shoud be. LH2 should be closer to LH1 as the canonical transcription fits the speech better. LH3 absorbes the overall acoustic characteristics of the speech sounds, thereby allowing LH1 to capture the phone specific acoustic characteristics.

LHs and LRs were calculated for each word individually, and were then used to calculate an average LR per utterance. For this purpose the word segmentations obtained with the forced Viterbi alignment are used. Forced Viterbi alignment has also been used in our previous research to calculate various temporal measures [4, 5], of which we will only use rate of speech in this paper:

* ros = # segments / total duration of speech plus pauses.

The automatic measures were calculated for each utterance. Next, an average score was calculated for all five utterance within a set. In this paper results of two likelihood ratios (LRs) are presented:

* LR1 = LH1/LH2
* LR2 = LH1/LH3

# 3. RESULTS

In this section the results of the present experiment are presented in the following order. In section 3.1. we report the results concerning the scores of pronunciation quality assigned by the three groups of experts. In 3.2. we analyze the results concerning the automatic measures of pronunciation quality. Finally, in 3.3 the correlations between these two types of results are considered.

## 3.1. Expert Ratings of Pronunciation Quality

The results concerning the scores assigned by the three groups of raters have been discussed in great detail in [4], where we reported on intrarater reliability, interrater reliability and, in particular, on the advantages of using standard scores. On the basis of this latter observation, the analyses to be presented in this paper are all carried out on standard scores. Table 1 shows the degree of interrater reliability for the standard scores of the three rater groups for the four scales.

| | **interrater reliability** | | | |
|---|---|---|---|---|
| | **OP** | **SQ** | **FL** | **SR** |
| **ph** | .98 | .98 | .96 | .91 |
| **st1** | .96 | .95 | .94 | .88 |
| **st2** | .96 | .93 | .92 | .91 |

**Table 1** Interrater reliability (Cronbach's α) for the three rater groups for the four scales.

After having established that the expert ratings are reliable and can be used for further analyses, we calculated the degree of correlation between the four scales for each rater group. The results are shown in Table 2.

| | | **OP** | **SQ** | **FL** | **SR** |
|---|---|---|---|---|---|
| **OP** | ph | | .97 | .87 | .73 |
| | st1 | | .96 | .87 | .60 |
| | st2 | | .91 | .77 | .64 |
| **SQ** | ph | | | .86 | .69 |
| | st1 | | | .91 | .61 |
| | st2 | | | .76 | .62 |
| **FL** | ph | | | | .87 |
| | st1 | | | | .83 |
| | st2 | | | | .83 |

**Table 2** Correlations between the four scales for the three rater groups.

As is clear from Table 2, the correlations between the four scales are very high for the three rater groups. However, there are small differences. For instance, the scale speech rate is clearly more highly correlated with fluency than with overall and segmental quality. This is not surprising if we consider that fluency and speech rate should represent temporal properties of speech, while

the other two scales should be more related to spectral properties. For each rater group the highest correlations are those of OP with SQ. Even though the ratings of OP and SQ were given on separate occassions the correlations are very high (varying from 0.91 to 0.97), thus suggesting that SQ is the most important factor for human ratings of pronunciation quality.

## 3.2. Automatic Pronunciation Measures

Several different likelihood ratios (LRs) were calculated. Only two of them, those with the highest correlations with the human ratings, are presented here. Before analyzing the relationship between the automatic measures and the expert pronunciation scores, it may be useful to investigate the relationships between the various automatic measures, as this may contribute to our understanding of the data.

| | **LR2** | **ros** |
|---|---|---|
| **LR1** | .96 | .63 |
| **LR2** | | .56 |

**Table 3** Correlations between the three automatic measures.

In Table 3 it can be seen that LR1 and LR2 are strongly correlated. Even though the procedures used to calculate these two likelihood ratios are quite different (see section 2.4), the resulting scores seem to be very similar. From Table 3 it also appears that both LR1 and LR2 have a fairly high correlation with ros.

## 3.3. Expert Pronunciation Scores and Automatic Measures

After having seen that the two log-likelihood ratio measures are correlated with rate of speech, we were very curious to see how they are related to the four types of expert judgements. These results are shown in Table 4 below.

| | | **OP** | **SQ** | **FL** | **SR** |
|---|---|---|---|---|---|
| **LR1** | ph | -.49 | -.45 | -.64 | -.64 |
| | st1 | -.54 | -.55 | -.68 | -.68 |
| | st2 | -.47 | -.44 | -.55 | -.62 |
| **LR2** | ph | -.42 | -.39 | -.59 | -.62 |
| | st1 | -.47 | -.49 | -.64 | -.65 |
| | st2 | -.38 | -.38 | -.48 | -.59 |
| **ros** | ph | .82 | .79 | .93 | .92 |
| | st1 | .83 | .79 | .91 | .89 |
| | st2 | .77 | .76 | .90 | .89 |

**Table 4**. Correlations between the automatic measures and the pronunciation scores by the three rater groups (ph, st1, st2).

It is clear from Table 4 that the four rating scales are much more strongly correlated with ros than with the two log-likelihood ratio measures. Apparently, ros is a better overall predictor of the human ratings than the LRs. Furthermore, it can be observed that the human scores are more strongly correlated with LR1 than with LR2. A possible explanation for this finding is that since the phone recognizer uses more acoustic models than the BPC recognizer, it can make a more detailed transcription of the signal, and thus has more discriminative power.

Apart from the correlations between the automatic and the human scores, we also performed stepwise multiple regression analyses in which OP was the criterion and ros was entered as first predictor. Only a slight increase in the multiple correlation coefficient was observed when LR1 or LR2 were entered in the regression equation after ros. Thus, it appears that the LRs do not contain information that is not present in the ros scores.

# 4. DISCUSSION AND CONCLUSIONS

This paper has two goals, viz. to investigate whether conventional CSRs can provide measures of pronunciation quality that can be used in automatic assessment and to improve our basis understanding of assessment of pronunciation quality. Our previous research has revealed that temporal measures are very good predictors of OP [4, 5]. For instance the correlations between ros and OP vary between 0.77 and 0.83. For various reasons measuring temporal variables alone will not suffice for automatic pronunciation assessment: 1. the speaker can speak the target sentence (very) fast, but with a poor pronunciation or 2. he/she can speak a different sentence (very) quickly. Thus, we need independent measures that will help to detect either one of these problem conditions. Even though the LRs used in the present research are strongly correlated with ros, we still believe that there are good reasons to assume that these measures will signal any situation in which the speaker produces the wrong sentence with approriate ros, instead of the prompted utterance.

The fact that LRs do not seem to contain additional information that is not already present in ros is due to the fact that speech rate and pronunciation quality in our data are very closely related. This is obvious from the correlations in Table 2. Two interpretations are possible at this point: 1. these two aspects are really so interrelated, and 2. this is a kind of 'artefact' of our data, which are limited to read speech. As a matter of fact, it is possible that the underlying construct 'proficiency' in read speech is reflected in good segmental quality AND fast speech rate at the same time. This would entail that for read speech data it is impossible to find a variable that is correlated with segmental quality and not with rate of speech. The only way to separate the variables would then be to search for ways of prompting utterances in which either rate or pronunciation quality is not at stake. Even if such prompting situations can be devised, it is questionable whether they will have any ecological validity. Thus, for the moment we have to bear with data in which the two aspects are intertwinded.

The question at this point is whether LRs can distinguish utterances that have globally correct phonemic make-up, but that are articulated with a 'foreign accent'. In general, LRs are best computed by comparing the LH of a model with that of a specific anti-model: LR = LH(model) / LH (anti-model) [8]. This is somewhat similar to the situation described in [3], where the task was to judge speech from a relatively homogeneous group (Native Americans) who have to speak Parisian French. In such a situation it may well be possible to train anti-models that target 'typical'pronunciation problems. Since our instrument has to

measure how well foreign speakers from a wide range of L1 backgrounds speak Dutch, we cannot take recourse to specific anti-models. Moreover, we did not want to limit the reference to Standard Dutch, but we wanted to allow all generally accepted regional Dutch accents. For this reason we used the Polyphone data base to train the CSR, because this data base contains all varieties of Dutch, regionally balanced.

A possible explanation of these results is that our anti-model may not be optimal and that an anti-model trained on non-native Dutch would work better. Unfortunately, at the moment we do not have such a data base which is large enough to train the anti-model. It is therefore possible that with a better anti-model we could obtain better results than those reported in this paper.

On the basis of these results it may be legitimate to wonder about what constitutes foreign accent. Which acoustic phonetic properties of a speech signal are responsible for the percept 'foreign accent'? Unfortunately, there is no simple answer to this question. In any case, the answer seems to depend very much on the combination of the first and second language. Some language pairs lead to characteristic insertions, deletions [9] or substitutions [10] in non-native speech. In principle, insertions and deletions can be detected by means of ASR techniques, e.g. by comparing forced decodings with and without the relevant segments in the transcriptions. However, substitutions may be much harder to detect. Distortions (i.e., segments that are produced in a recognizable way, but yet are different from the way natives pronounce them) are even more difficult to pin down in acoustic phonetic terms. It may very well be that non-native distortions can only be tracked down through very precise models of the temporal dynamics of co-articulation. It is well known that the present generation of HMMs do a very bad job in modeling temporal dynamics. In conclusion, it seems unlikely that word or utterance based likelihoods obtained with a conventional HMM recognizer can capture the acoustic phonetic details that are responsible for the perception of foreign accent. HMM recognizers may be deployed to score the number of inserted and deleted segments, at least as long as the type of segments that is affected can be accurately predicted from the knowledge of the first language of a learner of a specific second language.

To recapitulate, the results found so far show that a good prediction of OP can be obtained on the basis of ros and that LRs can be used to prevent that speakers who produce the wrong utterances with approriate ros get high pronunciation scores. However, it seems that further research is needed to determine whether a more refined assessment of 'foreign accent' is possible. To this end, one could look in different directions. For instance, it may be useful to look for measures that take more account of speech dynamics than those used so far. Alternatively, one could try to improve the calculation of log-likelihood ratios by looking for more appropriate anti-models. In any case it seems that this kind of research could profit from more insight into what 'foreign accent' really is, which is to say that further research is needed to determine what constitutes 'foreign accent'.

## 5. ACKNOWLEDGMENTS

This research was supported by the Ministry of Economic Affairs (SENTER), the Dutch National Institute for Educational Measurement (CITO), Swets and 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.

## 6. REFERENCES

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

2. Neumeyer L., Franco H., Weintraub M., and Price P. "Automatic text-independent pronunciation scoring of foreign language student speech", *Proc. ICSLP '96*, Philadelphia, 1457-1460, 1996.

3. Franco H., Neumeyer L., Kim Y., and Ronen O. "Automatic pronunciation scoring for language instruction", *Proc. ICASSP 1997*, München, 1471-1474, 1997.

4. Cucchiarini C., Strik H., and Boves L. "Automatic pronunciation grading for Dutch", *Proc. STiLL 98*, Marholmen, 95-98, 1998.

5. Cucchiarini C., Strik H., and Boves L. "Using speech recognition technology to assess foreign speakers pronunciation of Dutch", *Proc. New Sounds 97*, Klagenfurt, 61-68, 1997.

6. Strik, H., Russel, A., Van den Heuvel, H., Cucchiarini, C., Boves, L., "A spoken dialogue system for the Dutch public transport information service", *International Journal of Speech Technology*, 121-13, 1997.

7. den Os, E.A., Boogaart, T.I., Boves, L., Klabbers, E., "The Dutch Polyphone corpus", *Proc. Eurospeech 95*, Madrid, 825-828, 1995.

8. Lee, C.H. "A unified statistical hypothesis testing approach to speaker verification and verbal information verification utterance verification", *Proceedings COST workshop Rhodos*, 63-72, 1997.

9. Abrahamsson, N., "Vowel epenthesis of final /sC(C)/ clusters in Spanish speakers' L1 and L2 production: puzzle or evidence for natural phonology", *Proc. New Sounds 97*, Klagenfurt, 8-17, 1997.

10. Kerschofer-Puhalo, N., "Vowel substitutions in Geran as a foreing language", *Proc. New Sounds 97*, Klagenfurt, 167-175, 1997.