

# Automatic Pronunciation Grading For Dutch

*Catia Cucchiarini, Helmer Strik and Lou Boves*

*A<sup>2</sup>RT*, Dept. of Language & Speech, University of Nijmegen, The Netherlands  
{Cucchiarini, Strik, 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, special attention is paid to expert ratings of pronunciation, because they are used as a reference to validate the pronunciation scores obtained automatically. It is shown that the ratings can differ between raters and rater groups and it is concluded that these differences should be taken into consideration before going on to develop an automatic system for pronunciation grading.

## 1. Introduction

In the last few years, various attempts have been made at developing automatic methods for pronunciation scoring by using speech recognition technology [1, 2, 3, 4]. In general, the performance of such systems is evaluated by comparing the machine scores with pronunciation scores assigned by human experts. So far, high correlations have been reported between expert pronunciation ratings and various automatically obtained measures of speech quality. In particular, temporal measures of speech, such as segment duration scores and speaking rate [2, 4], turn out to be strongly correlated with expert pronunciation ratings. More recently, slightly higher correlations have been reported between human scores and HMM phone posterior probabilities [3].

It is obvious that in this kind of research the importance of the human ratings cannot be overestimated, because they are the reference which is used to validate the scores obtained automatically. Also in the study reported on in this paper human ratings are taken as reference to evaluate the performance of the speech recognizer. However, before making any decisions as to the further development of our system, we decided to gain more insight into the way in which pronunciation is evaluated by experts. First of all, we asked the experts to score different aspects of pronunciation quality, because it is known from the literature that expert ratings of pronunciation can be affected by different speech characteristics. Since we will calculate correlations between human ratings and machine scores, it is important to know exactly what the expert ratings represent. Second, we decided not to limit ourselves to one group of experts, because it is possible that the ratings assigned vary with the experts in question. Given that the expert ratings will further be used as our reference for validating the machine scores, it is important to make a well-motivated choice at the beginning. In this paper, we will not be so much concerned with the scores assigned by the experts to the

various aspects of pronunciation quality and with their relation to the machine scores, but we will focus on the differences observed between the scores of the different raters.

## 2. Method

### 2.1. Speakers

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. For further details, see [4].

### 2.2. Raters

Since in this experiment specific aspects of pronunciation quality had to be evaluated (see 2.4), raters with a high level of expertise were required. In selecting experts to assess non-native pronunciation of Dutch we could choose from among different groups. Phoneticians are obvious candidates, because they are experts on pronunciation in general. Teachers of Dutch as a second language would seem to be another obvious choice. However, it turned out that, in practice, pronunciation problems of people learning Dutch as a second language are usually not addressed by language teachers, but by specially trained speech therapists. In other words, speech therapists would seem to better qualify as 'non-native pronunciation experts' than language teachers. Finally, three groups of raters were selected. The first group consisted of three expert phoneticians (ph) with considerable experience in judging pronunciation and other speech and speaker characteristics. The second and the third groups each consisted of three speech therapists (st1 and st2) who had considerable experience in treating students of Dutch with pronunciation problems.

### 2.3. Speech material

Each speaker read two sets of five phonetically rich sentences (about one minute of speech per speaker) over the telephone. 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 (for further details, see [4]).

### 2.4. Expert ratings of pronunciation quality

The experts rated four different aspects of oral delivery: Overall Pronunciation (OP), Segmental Quality (SQ), Fluency (FI) and Speech Rate (SR). 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. Overall Pronunciation, 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. 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 [4].

### 2.5 Automatic pronunciation grading

A standard CSR system with phone-based HMMs was used to calculate automatic scores (for further details about the speech recognizer and the corpus used to train it, see [4]). Of all automatic measures that we calculated, here we will discuss those that are better correlated with the human ratings. These measures are all related to temporal characteristics of speech. The automatic scores were obtained for each set consisting of five sentences. In computing the automatic scores, a form of forced Viterbi alignment was applied. The following measures were calculated:

- td2 = total duration of speech plus pauses
- ptr = phonation time ratio (total duration of speech without pauses / td2)
- ros = rate of speech (# segments / td2)
- art = articulation rate (# segments / total duration of speech without pauses)

## 3. Results

Both for the automatic measures and for the expert ratings, speaker level scores were obtained by averaging the scores for the two sentence sets.

### 3.1 Expert ratings of pronunciation quality

Each rater scored 12 sentence sets twice, so that we could calculate intrarater reliability (see Table 1).

Table 1. Intrarater reliability (Cronbach's  $\alpha$ ) for the various scales (OP, SQ, FI and SR) and the raters in the three groups.

	Phoneticians			Speech therapists 1			Speech therapists 2		
	1st	2nd	3rd	1st	2nd	3rd	1st	2nd	3rd
OP	.97	.95	.99	.85	.94	.97	.93	.92	.98
SQ	.96	.98	.93	.86	.98	.99	.74	.94	.95
FI	.97	.94	.95	.94	.97	.96	.90	.76	.91
SR	.94	.76	.74	.73	.84	.88	.85	.94	.72

Except for a few instances, intrarater reliability is considerably high for the various raters and the various scales. Furthermore, interrater reliability was calculated on the basis of a 44-set overlap, i.e. 44 sentence sets that were scored by all three raters in each group. Since native speakers, and in particular standard language speakers, consistently receive higher scores than the non-native 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: C1. SDS NS NNS (all three groups of speakers), C2. NS NNS (without Standard Dutch speakers) and C3. NNS (only foreign speakers). As is clear from Table 2, interrater reliability is very high, even in the least favorable condition (C3).

Table 2. Interrater reliability ( $\alpha$ ) for three rater groups in three different conditions.

	Phoneticians			Speech therapists 1			Speech therapists 2		
	C1	C2	C3	C1	C2	C3	C1	C2	C3
OP	.97	.96	.89	.95	.93	.89	.95	.93	.87
SQ	.97	.97	.92	.95	.93	.85	.90	.84	.74
FI	.96	.95	.96	.93	.91	.88	.90	.88	.83
SR	.86	.84	.87	.82	.76	.81	.84	.82	.84

Subsequently, we checked the degree of correlation between the ratings assigned by the three rater groups. The results are presented in Table 3.

Table 3. Correlations between the ratings of the three rater groups (ph, st1, st2).

	OP	SQ	FI	SR
ph - st1	.92	.90	.94	.90
ph - st2	.80	.57	.82	.88
st1 - st2	.90	.69	.83	.81

It is known that measurement errors affect the size of the correlation coefficient, therefore, the correction for attenuation formula was applied, so as to allow comparisons between the various coefficients. As is clear from Table 3, the correlation coefficients differ for the various groups and the various scales. In order to find out how these differences came about, we analyzed the data in more detail. Besides considering interrater reliability, we also checked the degree of interrater agreement. Closer inspection of the data revealed that the means and standard deviations varied between the raters in a group, but also between the raters in different groups who rated the same speech material. The agreement within a group of raters has obvious consequences for the correlation coefficient computed

between the combined scores of the raters and another set of data (i.e. the ratings by another group or the machine scores). If the raters differ as to the absolute values of their ratings, the correlation coefficient between the combined scores and the other set of scores is going to be lower than it would be if the absolute values were similar. Furthermore, when several groups are compared, differences in correlation may be observed, which are a direct consequence of differences in the degree of agreement between the ratings. This is something that should be kept in mind when considering the correlations between the expert ratings and the machine scores.

### 3.2 Relation between expert ratings and automatic scores

The correlations (also corrected for attenuation) between the four automatic measures and the four rating scales for all three rater groups are presented in Table 4.

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

		Overall	Segm.	Fluency	Speech
		quality	quality		rate
td2	ph	-.73	-.68	-.90	-.82
	st1	-.78	-.77	-.97	-.86
	st2	-.72	-.65	-.86	-.85
ptr	ph	.69	.64	.83	.75
	st1	.76	.74	.92	.75
	st2	.70	.68	.85	.78
ros	ph	.76	.72	.92	.83
	st1	.80	.79	.93	.87
	st2	.75	.70	.85	.85
art	ph	.72	.68	.88	.80
	st1	.74	.73	.86	.88
	st2	.70	.63	.76	.81

As appears from Table 4, all automatic measures are strongly correlated with the expert ratings. Furthermore, since the automatic scores are based on temporal speech characteristics, they are also more strongly correlated with the human ratings related to speech timing, such as Fluency and Speech Rate, than to the other scales Overall and Segmental Quality.

Table 4 also reveals that the correlations between machine scores and expert ratings differ for the three groups of raters: the correlations are highest for the st1 group and lowest for the st2 group. On average the differences are about 0.05 between ph and st1 and about 0.06 between st1 and st2, while ph and st2 differ by only 0.01. These differences turn out to be significant according to analysis of variance ( $F_{2,30}=23.40$ ,  $p=.000$ ). However, since it may be questionable whether data of this kind should be subjected to analysis of variance, we also carried out a nonparametric test of significance for related samples, the Friedman test. In this case the

differences in correlation also turned out to be significant ( $\chi^2=17.56$ ,  $p=.0002$ ).

As we pointed out before, the differences in scores between the raters in each group could be responsible for these differences. Therefore, we decided to normalize for the differences in the values by using standard scores instead of raw scores. For this normalization we used the means and standard deviations of each rater in the overlap material, because in this case all raters scored the same samples. However, these values hardly differed from the means and standard deviations for the total material. Table 5 shows the correlation coefficients between the standard expert scores and the machine scores (also corrected for attenuation).

Table 5. Correlations between the automatic measures and the standard scores by the three rater groups (ph, st1, st2).

		Overall	Segm.	Fluency	Speech
		quality	quality		rate
td2	ph	-.79	-.75	-.91	-.90
	st1	-.81	-.77	-.94	-.88
	st2	-.73	-.70	-.91	-.88
ptr	ph	.76	.73	.86	.86
	st1	.78	.74	.88	.78
	st2	.72	.72	.89	.80
ros	ph	.82	.79	.93	.92
	st1	.83	.79	.91	.89
	st2	.77	.76	.90	.89
art	ph	.76	.73	.88	.86
	st1	.76	.73	.84	.88
	st2	.71	.68	.81	.86

If we compare Table 5 with Table 4 two things can be observed: the differences between the groups are smaller and the correlations are stronger. On average, the differences between the groups are now 0.03 between ph and st2, and between st1 and st2, while the difference between ph and st1 is much smaller (0.0025). However, these differences are still significant according to analysis of variance ( $F_{2,30}=12.8$ ,  $p=.000$ ) and the Friedman test ( $\chi^2=12.88$ ,  $p=.0016$ ). As to the increase in correlation, on average it is about 0.03, but it is different for the three rater groups: it is 0.056 for ph, 0.004 for st1 and 0.037 for st2. These results are in line with our expectations. Normalization leads to smaller differences in correlation between the rater groups and to higher correlations. Moreover, the gain in the size of the correlation coefficient is different for the three groups. Since the st1 group exhibited the smallest differences between the absolute values of the ratings, it is also the group for which normalization leads to the smallest improvement. The reverse applies to the other two groups.

If we now consider the correlations between the

normalized scores of the three rater groups (Table 6), we notice that these are considerably higher than those presented in Table 3. In other words, while the different degrees of agreement within the rater groups obscure the relationships between the groups, normalization contributes to clarifying these relationships. A clear understanding of how the ratings of the various groups relate to each other is necessary, because these correlations constitute some kind of upper limit for the correlations between human ratings and machine scores.

Table 6. Correlations between the ratings of the three rater groups (ph, st1, st2).

	OP	SQ	FI	SR
ph - st1	.96	.91	.94	.93
ph - st2	.90	.87	.90	.86
st1 - st2	.94	.84	.90	.89

#### 4. Discussion

The investigation reported on here was carried out within the framework of a study which aims at developing an automatic pronunciation scoring system for Dutch. In this paper we have considered how pronunciation ratings assigned by different groups of pronunciation experts are related to each other and to speech quality scores computed by an automatic speech recognizer. Special attention was paid to the ratings assigned by various groups of expert raters. The rationale behind investigating expert pronunciation ratings is that they are used as a reference in automatic pronunciation grading. Given the importance attached to expert ratings, it is interesting to know whether the choice of the experts can have consequences for the results obtained. Our findings show that although different raters in a group may achieve a high level of reliability as a group, they can still differ from each other in the way in which they use the rating scales, so that their mean ratings are different. In turn, this can affect the correlations computed between the combined scores of the raters in a group and those of other rater groups or those of the machine.

This is indeed what we observed in our data. To obviate this, we normalized the scores by calculating standard scores. In the correlations computed after normalization, different changes could be observed.

First of all, the correlations between the ratings of the three groups and the machine scores are more similar. Although the differences remain statistically significant, it does not seem that we can conclude, on the basis of these results, that the outcome of the validation procedure is strongly dependent on the choice of the expert rater group taken as a reference.

Second, as expected, almost all correlations between the rater scores and the machine scores are higher after normalization. The average increase in correlation is

about 0.03, which is comparable to the increase obtained by using posterior probabilities instead of duration scores [3]. So it seems that in addition to looking for alternative automatic measures that better correlate with the human ratings, one way of obtaining higher correlations is by normalizing the data for possible differences in the mean ratings of the experts.

Third, the correlations between the ratings of the three groups are higher and more similar. A comparison of these correlations with those between expert ratings and machine scores suggests that trying to increase predictive power does not make much sense, because the correlations between man and machine are very similar to those between experts. Therefore, our future work will not be directed so much at improving the predictive power of our measures, but rather at implementing automatic measures that are related to aspects of pronunciation other than the temporal one. This should prevent fast speakers with a poor pronunciation from getting high pronunciation scores.

#### 5. Conclusions

On the basis of the results presented above, it can be concluded that the choice of the rater expert group has a small impact on the results obtained. On the other hand, taking the differences between the scores assigned by different raters into account can contribute to achieving higher correlations between machine scores and expert ratings. In this way greater insight into the relationships between the scores assigned by different rater groups may also be gained.

#### Acknowledgments

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

#### References

- [1] Bernstein J, Cohen M, Murveit H, Rtschev D, and Weintraub M (1990). Automatic evaluation and training in English pronunciation, *Proc. ICSLP '90*, Kobe, 1185-1188.
- [2] Neumeyer L, Franco H, Weintraub M and Price P (1996). Automatic text-independent pronunciation scoring of foreign language student speech, *Proc. ICSLP '96*, Philadelphia, 1457-1460.
- [3] Franco H, Neumeyer L, Kim Y and Ronen O (1997). Automatic pronunciation scoring for language instruction. *Proc. ICASSP 1997*, München, 1471-1474.
- [4] Cucchiari C, Strik H, Boves L (1997). Using speech recognition technology to assess foreign speakers' pronunciation of Dutch, *Proc. New Sounds 97*, Klagenfurt, 61-68.