L2 PRONUNCIATION QUALITY IN READ AND SPONTANEOUS SPEECH

Helmer Strik, Catia Cucchiarini, Diana Binnenpoorte

A^{2}RT, Dept. of Language and Speech, University of Nijmegen, the Netherlands

{Strik, C.Cucchiarini, D.Binnenpoorte}@let.kun.nl; http://lands.let.kun.nl/

ABSTRACT

This paper describes two experiments aimed at exploring the relationship between objective properties of speech and perceived pronunciation quality in read and spontaneous speech, with a view to determining whether such quantitative measures can be used to develop objective pronunciation tests. Read and spontaneous speech of two groups of 60 learners of Dutch as a second language was scored for pronunciation quality by human raters and was analyzed by means of a continuous speech recognizer to calculate six quantitative measures of speech quality related to speech timing. The results show that quantitative, temporal measures of speech are strongly related to pronunciation quality, in both read and spontaneous speech, although not all variables suitable for measuring pronunciation quality in read speech are as effective in spontaneous speech.

1. INTRODUCTION

Recent attempts at developing automatic methods for pronunciation scoring by using continuous speech recognizers (CSRs) have revealed that automatically obtained measures of speech quality are strongly correlated with pronunciation scores assigned by human experts. However, since most of these studies concern read speech, it is legitimate to question whether these results would hold for speech which is not read, such as extemporaneous or spontaneous speech.

In an attempt to find an answer to this question we decided to carry out experiments with read and spontaneous speech and compare the results. These two experiments will be referred to as Experiment 1 (read speech) and Experiment 2 (spontaneous speech). In Experiment 1 we investigated speech of 20 natives and 60 non-natives. Although this experiment has already been presented in detail in [2, 3], the data concerning the 60 non-native speakers were not presented so explicitly as they are in this paper. In any case, here we will limit ourselves to providing only the Experiment 1 data and details that are necessary to make comparisons between read speech (Experiment 1) and spontaneous speech (Experiment 2) of learners of Dutch as a second language (DSL). More details about this comparison can be found in [4].

In both experiments, a dual approach was adopted, i.e. the speech material was evaluated by expert raters and by a CSR. The aim of the present paper is to explore the relationship between objective properties of speech and perceived pronunciation quality in read and spontaneous speech, with a view to determining whether such objective measures can be used to develop objective pronunciation tests.

2. METHOD

2.1. Speakers and speech material

Experiment 1

The speakers involved in this experiment are 60 learners of DSL. Three proficiency levels were distinguished: PL1 = beginner, PL2 = intermediate, and PL3 = advanced. Each speaker read two sets of 5 phonetically rich sentences (about one minute of speech per speaker) over the telephone. An elaborated orthographic transcription of all the speech material was made before the latter was used for the experiment. For further details, see [2, 3].

Experiment 2

For this experiment we selected an existing test, the *Profieltoets*, developed by the Dutch National Institute for Educational Measurement (CITO). In the speaking component of this test the subjects answer a number of questions extemporaneously, so that they produce spontaneous speech. Among the candidates who took part in the *Profieltoets* in June 1998, 60 subjects of two different proficiency levels were selected: a lower proficiency group (LP) at the beginner level, and a higher proficiency group (HP) at the intermediate level.

An important requirement in selecting the items was that they elicit relatively long answers, which is a necessary condition for assessing aspects such as fluency and speech rate and for calculating some of the machine temporal measures. On average, each of the 30 LP subjects effectively talked for about 70 s for all eight items, while each of the 30 HP subjects talked for 170 s in total.

The two subject groups performed different tasks. The LP items contain questions that can be answered immediately by the candidate without much thinking: a situation is presented and the candidate has to indicate what he/she would say in that context. The HP items, on the other hand, contain questions that require more preparation to be answered, i.e. the candidate has to choose between various possibilities and has to explain why he/she made that choice. In other words, the HP group carried out cognitively more demanding tasks than the LP group and this difference could have an effect on the results.

The speech material was recorded in language laboratories onto audio cassettes and then digitized. Before it was analyzed by the CSR, an elaborate orthographic transcription was made [4].

2.2. Expert ratings

All raters in both experiments evaluated four different aspects of pronunciation quality: Overall Pronunciation (OP), Segmental Quality (SQ), Fluency (FL) and Speech Rate (SR). All raters listened to the speech material and assigned scores individually.

	read speech								spontaneous speech			
	PL1		PL2		PL3		all NNS		LP		HP	
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
OP	4.32	1.13	4.22	1.34	5.30	1.15	4.65	1.32	5.79	0.91	4.72	1.03
SQ	4.18	1.32	4.33	1.24	5.46	0.97	4.74	1.27	5.37	0.90	4.41	0.98
FL	4.65	2.01	5.00	1.81	7.36	0.95	5.85	1.96	5.64	0.88	4.80	1.06
SR	-1.37	1.61	-1.07	1.33	0.43	0.68	-0.55	1.40	1.15	0.98	0.29	1.08

Table 1: Means and standard deviations for read and spontaneous speech of speakers of different proficiency levels.

They could listen to the speech fragments as often as they wanted. 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.

Experiment 1

Three groups of experienced raters were selected: 1. three phoneticians (ph), 2. three speech therapists (st1), and 3. a second group of three speech therapists (st2). The scores were not assigned to each individual sentence, but to each set of five phonetically rich sentences. Next, the average score per subject was calculated. For further details, see [2, 3].

Experiment 2

The scoring sessions for Experiment 2 were organized by CITO according to the procedure that is usually followed for the *Profieltoets*. Ten teachers of DSL were employed as raters. A group of five teachers evaluated the 30 LP speakers and another group of five teachers evaluated the 30 HP speakers. There was no overlap of speakers between the two rater groups. Each of the five raters assigned one score for each of the four scales per set of eight items for one speaker.

Experiment 1 vs. Experiment 2

Two essential differences between the two experiments should be mentioned. First, in Experiment 2 two different groups of raters were assigned to the two groups of speakers, whereas in Experiment 1 the same group of raters evaluated all speakers. This point should be borne in mind because it has consequences for the analyses that can be carried out and for the results of these analyses.

Second, the phoneticians and speech therapists involved in Experiment 1 simply judged the speech of a number of speakers without having information on the proficiency level of each speaker, except the cues that they could derive from the speech itself. The language teachers in Experiment 2, on the other hand, were judging candidates in an examination and therefore knew whether a speaker was in the beginner or intermediate group.

2.3. Automatic pronunciation grading

A standard CSR system with phone-based hidden Markov models was used to calculate automatic scores (for further details about the speech recognizer and the corpus used to train it, see [2, 3]). Of all automatic measures that we calculated, here we will only discuss the 6 measures that correlate best with the human ratings:

1. *ros* (rate of speech) = #phones/ tdur2
2. *ptr* (phonation/time ratio) = 100% * tdur1/tdur2
3. *art* (articulation rate) = #phones/tdur2
4. *#ps* (#pauses per unit time) = #pauses/tdur2
5. *mlp* (mean length of pauses) = mean length of all pauses
6. *mlr* (mean length of runs) = average #phones between pauses

In these measures a pause is defined as a silence of at least 0.2 s,
tdur1 as ‘total duration of speech without silences’, and
tdur2 as ‘total duration of speech with internal silences’.

These 6 measures are all related to temporal characteristics of speech. In Experiment 1 the automatic scores were obtained for each set consisting of five sentences and were then averaged over the two sets, while in Experiment 2 these scores were obtained per set of eight items.

3. RESULTS

3.1. Expert ratings

The expert ratings were analyzed to determine interrater reliability. This appeared to be relatively high, as it varied between 0.96 and 0.81 for read speech and between 0.89 and 0.80 for spontaneous speech.

We then calculated the mean and standard deviations for of all ratings in the two experiments. In Table 1 we can clearly see that the read speech scores, in general, gradually increase as we go from PL1 to PL3, which means that the more proficient speakers receive higher scores for all four scales. In the spontaneous speech data this relationship between proficiency and human pronunciation ratings does not seem to exist, as the scores for the HP speakers are lower than those for the LP speakers. Although one might argue that the scores for the two speaker groups are not really comparable because they were assigned by two different groups of raters, it seems that these results might be related to the context within which the evaluation was carried out. As explained above, the raters in Experiment 1 had no information about the proficiency level of each speaker, except the cues contained in their speech, whereas the raters in Experiment 2 knew to which proficiency group the speaker belonged. As a consequence, they probably judged pronunciation quality in relation to each speaker’s proficiency level, thus assigning higher scores to absolutely less proficient speakers if the desired level of pronunciation quality was relatively high, i.e. in the LP group. Another possibility is that these results are due to the differences between the tasks carried out by the LP and the HP group. The analyses of the objective pronunciation measures may shed light on this point.

	read speech								Spontaneous speech					
	PL1		PL2		PL3		all NNS		LP		HP		LP-HP	
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	Sd	$\bar{x}$	sd
ros	8.54	1.88	8.95	1.87	11.03	1.16	9.68	1.94	5.99	0.96	5.31	1.17	5.65	1.12
ptr	77.97	7.69	79.62	8.68	88.28	5.42	82.7	8.57	49.32	8.71	44.92	9.51	47.10	9.32
art	10.87	1.41	11.15	1.38	12.47	0.82	11.6	1.37	12.25	1.25	11.85	0.81	12.00	1.06
#ps	0.37	0.14	0.34	0.16	0.17	0.11	0.28	0.16	0.52	0.09	0.52	0.08	0.52	0.09
mlp	0.40	0.08	0.40	0.12	0.34	0.16	0.38	0.13	0.92	0.20	1.02	0.28	0.97	0.25
mlr	16.51	7.67	18.10	7.44	27.73	7.13	21.5	8.77	9.50	2.22	9.33	2.27	9.41	2.23

Table 2: Means and standard deviations for the seven quantitative measures for read speech and spontaneous speech of speakers of different proficiency levels.

3.2. Machine pronunciation assessment

In this section we analyze the quantitative variables in various respects. First, we calculate the mean and standard deviation for all variables for all groups. These results are given in Table 2. This table shows how the values for the different variables vary as a function of speech modality (read vs. spontaneous) and proficiency level. In order to see how the objective measures vary as a function of speech modality we can compare the means for read speech (column 8) with those pertaining to spontaneous speech (column 14). These comparisons indicate that for almost all variables the values drastically change as we go from read speech to spontaneous speech. In particular, *ros*, *ptr* and *mlr* are approximately halved, *#ps* is almost doubled, while *mlp* is almost tripled. *art*, on the other hand, hardly changes. In other words, these data suggest that, at least for non-native speakers, the differences between read and spontaneous speech are more related to the frequency and the length of pauses, rather than to the rate at which sounds are articulated. As a consequence, all measures in which pause frequency and pause length play a part, vary substantially between the two speech modalities.

In order to see how the quantitative measures vary as a function of proficiency level, we can compare columns 2, 4 and 6 within read speech and columns 10 and 12 within spontaneous speech. In the read speech material we observe gradual changes as we move from PL1 to PL3. The change is either an increase or a decrease, depending on the variable in question, but all changes indicate that the less proficient speakers also obtain lower scores in terms of the quantitative measures. In the spontaneous speech material the opposite seems to hold: the measures for the less proficient speakers indicate better pronunciation quality than those of the more proficient speakers. This is all the more remarkable because it holds for all measures. On the one hand, these findings are in line with those presented in the previous section: also in the human ratings the LP speakers were perceived as having better pronunciation quality than the HP speakers. On the other hand, these findings are contrary to our expectations and to the results concerning read speech. However, these results may seem less surprising if we consider in more detail the speech material used in Experiment 2, as will be explained in the Discussion section.

3.3. Relation between expert ratings and automatic scores

In this section we compare the automatically calculated measures of speech quality with the pronunciation scores assigned by the raters, in order to determine how and to what extent (temporal) quantitative properties of speech are related to perceived pronunciation quality in read and spontaneous speech. To this end the correlations between the two sets of scores in each experiment were calculated. For Experiment 1 we calculated the means over the scores assigned by the three rater groups, because the ratings of the three groups appeared to be very strongly correlated with each other [3]. For Experiment 2, on the other hand, the ratings assigned to the two groups of speakers are not directly comparable, because they were assigned by different raters and for different tasks. Consequently, the correlations were calculated for each group of speakers separately. In this way the variation in proficiency level, which was already lower in Experiment 2 as compared to Experiment 1, is further reduced with obvious consequences for the correlations.

Table 3 shows the correlations between the six automatic measures and the four rating scales for three different groups: a) read speech of DSL learners of different proficiency levels (RS), b) spontaneous speech of DSL learners with a lower proficiency level (SSLP), and c) spontaneous speech of DSL learners with a higher proficiency level (SSHP).

		OP	SQ	FL	SR
ros	RS	.75	.70	.92	.91
	SSLP	.46	.47	.57	.57
	SSHP	.33	.22	.39	.60
ptr	RS	.73	.69	.86	.79
	SSLP	.39	.40	.46	.47
	SSHP	.39	.26	.39	.53
art	RS	.64	.60	.83	.89
	SSLP	.00	.00	.06	.05
	SSHP	-.15	-.11	.05	.23
#ps	RS	-.70	-.67	-.85	-.74
	SSLP	-.40	-.43	-.33	-.39
	SSHP	-.30	-.35	-.49	-.41
mlp	RS	-.54	-.50	-.53	-.46
	SSLP	.03	.06	-.08	-.03
	SSHP	-.09	.03	.00	-.13
mlr	RS	.72	.69	.85	.76
	SSLP	.49	53	.49	.57
	SSHP	.50	.42	.65	.80

Table 3: Correlations between the automatic measures and the pronunciation ratings for the three groups (RS, SSLP, SSHP).

As appears from Table 3, the correlations for the read speech material are all higher than those for spontaneous speech, which was to be expected given the greater homogeneity of the samples in Experiment 2 with respect to proficiency level. Another result that was to be expected is that the automatic measures would be more strongly correlated with the human ratings related to speech timing, such as FL and SR, than to the other scales OP and SQ. This appears to be indeed the case, but the differences are very small and it is actually surprising that these quantitative temporal measures are such good predictors of pronunciation quality in general. Other things to be observed in this table are that *art* and *mlp* have weak correlations with the human ratings in the spontaneous speech experiment, while they exhibited strong (*art*) and reasonable (*mlp*) correlations in the read speech experiment. These results will be discussed in the following section.

DISCUSSION

In this paper we have presented two experiments on non-native pronunciation quality assessment in read and spontaneous speech in which a dual approach was adopted: pronunciation ratings assigned by experts to read and spontaneous speech produced by learners of DSL were compared with a number of quantitative measures that were automatically calculated for the same speech fragments. The data analyzed here provide interesting results.

First, these results reveal how the nature of the task carried out by the speaker affects the pronunciation scores, both those assigned by human raters and those obtained on the basis of quantitative measures. As mentioned above, the HP items require more cognitive effort than the LP items. In turn, this could explain the lower pronunciation scores since more cognitively demanding tasks are associated with a lower articulation rate and a lower phonation/time ratio [5, 6] and this is exactly what appears from the comparison between LP and HP in Table 2.

Second, with respect to the role played by the various quantitative variables these results show that it may vary depending on the speech modality and the specific task used to elicit the material. Table 3 reveals that for read speech the pronunciation ratings are strongly correlated with *ros*, *art*, *ptr*, *#ps* and *mlr*, while *mlp* has a less strong correlation. As pointed out in [2] this suggests that for perceived fluency, and here we see that is also holds for pronunciation quality in general, the frequency of pauses is more relevant than their average length. These findings are in line with those of previous investigations [1] and are corroborated by the data concerning the three proficiency levels in the read speech experiment: Table 2 shows that the differences between the proficiency levels with respect to *mlp* are relatively smaller than those concerning *#ps*. As already noted in [2] these results suggest that two factors are particularly important for perceived fluency in read speech: the rate at which speakers articulate the sounds and the frequency with which they pause.

With regard to spontaneous speech, Table 2 shows that the pronunciation ratings are relatively strongly correlated with *ros*, *ptr*, *#ps*, and *mlr*, while low correlations were found for *art* and *mlp*. It is clear that pauses are much more frequent in spontaneous speech than in read speech (see Table 2). This might explain why a variable that takes no account of pauses whatsoever, like *art*, has almost no relation with perceived pronunciation quality. Furthermore, if we consider the nature of all these variables we then have to conclude that pronunciation ratings of spontaneous speech are particularly related to variables that contain information about the frequency of the pauses, and these are *ros*, *ptr*, *#ps*, and *mlr*, but not *art* and *mlp*. In turn, this suggests that of the two factors that are strongly related to perceived fluency in read speech, namely the rate at which speakers articulate the sounds and the frequency with which they pause, the latter is most important for perceived pronunciation quality in spontaneous speech.

In addition, we can see in Table 3 that *mlr* is a better predictor of pronunciation quality in spontaneous speech than all other measures that do take pause frequency into account. What distinguishes *mlr* from the other measures is that *mlr* takes account not only of the frequency of the pauses but, to a certain extent, of their distribution: pauses are tolerated provided that sufficiently long uninterrupted stretches of speech are produced. Moreover, the predictive power of *mlr* is greater for SSHP, that is for speech material where the speaker has to present his/her arguments in a coherent and more organized manner and where the distribution of pauses is of course more important.

To conclude, the fact that such easy to compute temporal measures of speech appear to be so strongly correlated with perceived pronunciation quality suggests that the techniques employed in this research can be used for the purpose of automatic pronunciation assessment.

ACKNOWLEDGEMENTS

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 Zeitling and KPN Telecom.

REFERENCES

1. Chambers, F. “What Do We Mean by Fluency?” *System*, 4, 535-544, 1997.

2. Cucchiarini, C., Strik, H. and Boves, L. “Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology”, *J. Acoustic. Soc. Amer.*, Vol. 107 (2), 2000, pp 989-999.

3. Cucchiarini, C., Strik, H. and Boves, L. “Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms”, *Speech Communication*, 30 (2-s3), 2000, pp 109-119.

4. Cucchiarini, C., Strik, H., Binnenpoorte, D. and Boves, L “Towards an Automatic Oral Proficiency Test for Dutch as a Second Language: Automatic Pronunciation Assessment in Read and Spontaneous Speech” Proc. InStiLL, Dundee, August 2000.

5. Goldman-Eisler, F. *Psycholinguistics: Experiments in Spontaneous Speech*, Academic, New York, 1968.

6. Grosjean, F. “Temporal Variables Within and Between Languages, in Towards a Cross-Linguistic Assessment of Speech Production”, in H.W. Dechert and M. Raupach (eds.): Lang, Frankfurt, 39-53, 1980.