PRONUNCIATION EVALUATION IN READ AND SPONTANEOUS SPEECH: A COMPARISON BETWEEN HUMAN RATINGS AND AUTOMATIC SCORES

Catia Cucchiarini, Helmer Strik, Diana Binnenpoorte and Lou Boves
A2RT, Dept. of Language & Speech, University of Nijmegen, The Netherlands
{Cucchiarini, Strik, Binnenpoorte, Boves}@let.kun.nl, http://lands.let.kun.nl/staff/

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 analysed 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. In particular, measures that express the rate at which sounds are produced without taking the frequency and distribution of pauses into account appear to be unsuitable for measuring pronunciation quality in spontaneous speech.

1. Introduction

Recent attempts at developing automatic methods for pronunciation scoring by using automatic speech recognition (ASR) technology (Bernstein et al. 1990: 1185-1188, Neumeyer et al. 1996: 1457-1460, Franco et al. 1997: 1471-1474 and Cucchiarini et al. 1997: 61-68) have revealed that automatically obtained measures of speech quality are strongly correlated with pronunciation scores assigned by human experts. These studies provide interesting information not only about the possibilities of automatically scoring pronunciation, but also about the nature of the human scoring behaviour and its relation to machine scoring.

Unfortunately, most of these studies concern read speech, because this is the type of speech considered to be most amenable to automatic pronunciation scoring, given the state of the art in ASR technology. It is therefore legitimate to question whether these results would hold for speech which is not read, such as extemporaneous and spontaneous speech. It would be interesting to know, for instance, whether the same quantitative measures that were found to be strongly correlated with pronunciation quality in read speech would be equally important for pronunciation in spontaneous speech and/or, the other way round, whether there are measures that are suitable for spontaneous speech, but not for read speech. In an attempt to find an answer to this question we decided to carry out an experiment with spontaneous speech.

As explained in (Cucchiarini et al. 2000a: 109-119) our interest in this type of research is not only related to the possibilities of getting more insight into human pronunciation scoring, but also to the potential that this kind of research might have for the development of objective testing instruments for pronunciation grading, specially in the context of second language teaching and testing. Against this background we thought that it would be more advantageous to use an existing test of second language proficiency rather than collecting speech material especially for this experiment. In this way the material under study would be less of the 'laboratory' type and would be more similar to what is generally found in the 'field'. On the one hand this might have the disadvantage that the experimenter cannot control all aspects of the experiment. On the other hand, it has the considerable advantage that in this way external validity is achieved, since we are convinced that the importance of external validity cannot be overestimated in these kinds of studies and that the advantages of using a real test evaluated by real raters outweigh the disadvantages of using a less elegant experimental design. Therefore we looked for an already existing test of second language proficiency that would be suited for our purpose.

The test that was eventually selected for this experiment is the Profieltoets (Cito, Arnhem 1998). This is a test which was developed by the Dutch National Institute for Educational Measurement (Cito). In this test various skills are tested, but we limited our experiment to the subtest for speaking. This test is administered in a language lab to a group of several candidates simultaneously. The candidates have to answer questions which elicit unprepared answers. The speech can therefore be classified as extemporaneous and spontaneous speech. As in the experiment in (Cucchiarini et al. 2000a: 109-119), a dual approach was adopted in which the speech material was evaluated by a group of raters and by an automatic continuous speech recognizer (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. To pursue this aim we compare the data of the read speech experiment with those of the spontaneous speech experiment. These two experiments will be referred to as Experiment 1 (read speech) and Experiment 2 (spontaneous speech). In Experiment 1 we investigated speech of both natives and non-natives. Although this experiment has already been presented in detail in (Cucchiarini et al. 2000a: 109-119), the data concerning the 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 non-native speakers of Dutch, i.e. learners of Dutch as a second language (DSL).

2. Method

2.1 Speakers

2.1.1 Experiment 1

The speakers involved in this experiment are 60 non-native speakers (NNS) who all lived in The Netherlands and were attending or had attended courses in Dutch as a second language. They were selected to obtain a group that was sufficiently varied with respect to mother tongue, proficiency level and gender. Three proficiency levels were distinguished: PL1 = beginner, PL2 = intermediate and PL3 = advanced. For more detailed information on the composition of this sample, see (Cucchiarini et al. 2000a: 109-119 and Cucchiarini et al. 2000b: 989-999).

2.1.2 Experiment 2

The speakers involved in this experiment constitute a subgroup of the candidates who took part in the test Profieltoets in June 1998. In this investigation we analyzed the answers of 60 subjects of two differing proficiency levels: a lower proficiency group (BP) at the beginner level and a higher proficiency group (IP) at the intermediate level. Cito workers selected for us two subgroups of 30 speakers per proficiency level who varied with respect to gender and mother tongue.

2.2 Raters

2.2.1 Experiment 1

As explained in (Cucchiarini et al. 2000a: 109-119) in this experiment raters with a high level of expertise were employed because specific aspects of pronunciation quality had to be evaluated (see below). 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 group consisted of three speech therapists (st1 and st2) who had considerable experience in treating students of Dutch with pronunciation problems.

2.2.2 Experiment 2

In this experiment ten teachers of Dutch as a second language (DSL) were employed because they are normally used as raters for this kind of examination by Cito. To be able to work as raters for Cito these teachers have to follow a three-day course which they have to conclude with an examination.

The scoring sessions were organized by Cito according to the procedure that is usually followed for the Profieltoets. A group of five teachers evaluated the BP speakers and another group of five teachers evaluated the IP speakers. There was no overlap of speakers between the two rater groups.

2.3 Speech material

2.3.1 Experiment 1

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. An elaborated orthographic transcription of all the speech material was made before the latter was used for the experiment (for further details, see (Cucchiarini et al. 2000a: 109-119 and Cucchiarini et al. 2000b: 989-999)).

2.3.2 Experiment 2

The speech material used in this experiment consists of the answers given by the above-mentioned candidates to part of the items which constitute the Profieltoets. The test is available in two different versions for the two proficiency groups of beginner and intermediate. For this experiment eight items were selected for each version of the test. The items differed for the two proficiency groups, which is a consequence of choosing an existing test, because in this case we have less influence on the selection of the material. An important requirement in selecting the items was that they had to 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.

For the IP group we chose the so-called long tasks, in which the candidates have 30 s to answer each question. In these items the candidates have to answer questions and have to motivate choices among various possibilities.
The BP version of the test does not contain the long tasks, but only the short tasks, in which the subjects have 15 s at their disposal to answer each question. In these items a given situation is presented and the candidates have to indicate what they would say in that context. Among these tasks we chose those which, given the nature of the questions, would elicit reasonably long answers of at least a few words. For all items, the BP subjects effectively talked for about 58 s in total on average, while for the IP subjects the average was 75 s in total.

The speech material of Experiment 2 was recorded in language laboratories onto audio cassettes and was subsequently digitized. In this case the recording conditions were rather adverse: the subjects, who were taking an exam, were all sitting in one room and started to answer the questions almost at the same time, so that there was a lot of

background speech. Of this material also an elaborate orthographic transcription was made before the material was analysed by the CSR.

2.4 Expert ratings of pronunciation quality

All raters in both Experiment 1 and Experiment 2 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. 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.

2.4.1 Experiment 1

The scores were not assigned to each individual sentence, but to each set of five phonetically rich sentences. No specific instructions were given as to how to use the scales. However, before starting with the evaluation proper, each rater listened to five sets of sentences spoken by five different speakers, which were intended to familiarize the raters with the task they had to carry out and to help them anchor their ratings. As a matter of fact, the five speakers were chosen so as to give an indication of the range that the raters could possibly expect. 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 speakers were proportionally assigned to the three raters in each group. For further detail on this point, see (Cucchiarini et al. 2000a: 109-119 and Cucchiarini et al. 2000b: 989-999). The scores assigned by the three raters were then combined to compute correlations with the machine scores.

2.4.2 Experiment 2

Each of the five raters assigned one score per set of one speaker for each of the four scales. As in the experiment in (Cucchiarini et al. 2000a: 109-119), no specific instructions were given for pronunciation assessment, however these raters had all received a three-day training before starting to work as raters for Cito.

2.4.3 Experiment 1 versus 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 user group. As a consequence, they might have judged pronunciation in relation to each speaker's proficiency level, so that the same score would not have the same meaning in the two groups, but would represent better pronunciation quality in the IP group than in the BP group.

2.5 Automatic pronunciation gradings

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 (Cucchiarini et al. 2000a: 109-119 and Cucchiarini et al. 2000b: 989-999)). Of all automatic measures that we calculated, here we will discuss those that are best correlated with the human ratings. These 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.

In computing the automatic scores, a form of forced Viterbi alignment was applied. The following measures were calculated:

ros = rate of speech = # phones / total duration of speech including answer-internal pauses
ptr = phonation/time ratio = 100 % x total duration of speech without pause / total duration of speech including answer internal pauses
art = articulation rate = # phones / total duration of speech without pauses
#ps = # of silent pauses per unit time = # of answer-internal pauses of no less than 0.2 s / total duration of speech including answer-internal pauses
mlp = mean length of pauses = mean length of all answer-internal pauses of no less than 0.2 s
mlr = mean length of runs = average number of phones occurring between unfilled pauses of no less than 0.2 s

3. Results

In presenting the results of the two experiments, we will first pay attention to the ratings assigned by the various groups of raters on the basis of the four scales. Subsequently, the results concerning the objective measures of pronunciation quality will be examined. Finally, the relationship between the human-assigned ratings and the objective measures will be considered.

3.1 Expert ratings of pronunciation quality

The ratings assigned by the various rater groups involved in the two experiments, ph, st1 and st2 for Experiment 1 and RBP (raters for the BP group) and RIP (raters for the IP group) for Experiment 2, were analyzed to determine interrater reliability. The results of these analyses are shown in Table 1.

	OP	SQ	FL	SR
ph	.89	.92	.96	.87
st1	.89	.85	.88	.81
st2	.87	.74	.83	.84
RBP	.89	.82	.86	.89
RIP	.84	.81	.82	.80

Table 1. Interrater reliability coefficients (Cronbach's α) for the five rater groups and the four scales

As is clear from Table 1, the values for interrater reliability in Experiment 2 are comparable to those in Experiment 1. This may be surprising if we consider that the speech used in Experiment 2 was highly variable for each speaker with respect to syntax and vocabulary and that this kind of variation is known to affect ratings of speech quality such as fluency ratings (Riggenbach 1991: 423-441 and Freed 1995: 123-148). The relatively high reliability coefficients that were found in Experiment 2 may be ascribed to the fact that the raters involved in this experiment did receive training before starting their activities as raters at Cito.

Besides considering interrater reliability, we also checked the degree of interrater agreement. Closer inspection of the data revealed that in both experiments the means and standard deviations varied between the various raters. In other words, in both experiments the raters differed from each other in degree of strictness. Therefore, we decided to normalize for the differences in the values by using standard scores instead of raw scores. Further details on the normalization procedure applied in Experiment 1 can be found in (Cucchiarini et al. 2000a: 109-119). In Experiment 2 normalizing the scores was more straightforward, because all five raters in one group rated all speakers. For each rater we then subtracted his/her mean from each of his/her scores and the resulting scores were then divided by the standard deviation for that rater.

	read speech								spontaneous speech			
	PL1		PL2		PL3		all NNS		BP		IP	
	$\bar{x}$	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 2. Means and standard deviations for the raw scores for read and spontaneous speech of speakers of different proficiency levels

Table 2 shows the mean and standard deviations (raw scores) of the human ratings for the speakers in the two experiments. In Table 2 we can clearly see that the read speech scores vary for the three proficiency levels PL1, PL2 and PL3 and that, in general, they 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 IP speakers are lower than those for the BP 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 might have judged pronunciation quality in relation to each speaker's proficiency level, thus assigning higher scores to less proficient speakers if the desired level of pronunciation quality was lower, i.e. in the BP group. The analyses of the objective pronunciation measures may shed light on this point.

		SQ	FL	SR
OP	RS	.90	.78	.67
	BP	.97	.91	.88
	IP	.94	.89	.78
SQ	RS		.78	.61
	BP		.92	.89
	IP		.89	.78
FL	RS			.88
	BP			.95
	IP			.91

Table 3. Correlations among the different scales for read speech (RS) and spontaneous speech of speakers in the lower proficiency (BP) and the higher proficiency (IP) group

To get more insight into the human scoring of pronunciation quality in read and spontaneous speech, we analyzed the correlations among the various scales in both experiments. For Experiment 1 we calculated the average scores over the three rater groups, because these appeared to be strongly correlated with each other (Cucchiarini et al. 2000a: 109-119). We then computed the correlations among these average scores for all non-native speakers (RS).

As is clear from Table 3, all four scales are strongly correlated with each other, but there are differences. In particular, OP and SQ are more strongly correlated with each other than all other scales. FL and SR are also strongly correlated with each other, which is obvious given that both refer to temporal aspects of pronunciation quality. FL is the only scale that shows similarly strong correlations with the other three. This structure emerges for all three groups, RS, BP and IP.

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 4. 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).

	read speech								spontaneous speech					
	PL1		PL2		PL3		all NNS		BP		IP		BP & IP	
	$\bar{x}$	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 4. Means and standard deviations for the seven quantitative measures for read speech and spontaneous speech of speakers of different proficiency levels

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 almost 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 BP speakers were perceived as having better pronunciation quality than the IP 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 against the backdrop of what we mentioned above with respect to 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 (Cucchiarini et al. 2000a: 109-119). 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 to different kinds of speech. 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.

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

Table 5. Correlations between the automatic measures and the pronunciation ratings for the three groups (RS, SSBP, SSIP)

Table 5 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 (SSBP), and c) spontaneous speech of DSL learners with a higher proficiency level (SSIP).

As appears from Table 5, 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 almost no correlation 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.

4. 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.

These studies have revealed that it is possible to obtain reliable expert ratings of pronunciation quality both in read and spontaneous speech: reliability was reasonably high for all rater groups in both experiments (Cronbach's α varied between .74 and .96). These results may be surprising in view of the much lower degrees of reliability obtained in previous studies (Riggenbach 1991: 423-441 and Freed 1995: 123-148) and require some explanation. Various factors may have led to such high reliability coefficients in the two experiments. In Experiment 1 the raters did not receive specific instructions on how to use the evaluation scales, however they were highly trained and had received some indications concerning the proficiency levels that they could possibly expect before the evaluation proper started. In addition, since they evaluated read speech they could more easily concentrate on the speakers' pronunciation without being distracted by other variables such as syntax and vocabulary which were kept constant. In Experiment 2 this was not the case since each speaker gave different answers. However, also in this case the raters were highly trained and experienced. They had received training before starting their activities as raters and had participated in various rating sessions at Cito.

With respect to the major goal of this study, getting more insight into the nature of the human pronunciation scoring behavior and its relation to machine scoring in read and spontaneous speech, the data analysed here provide interesting results.

First of all, the results obtained in this study have shown that the various aspects of pronunciation quality investigated here have the same interrelations in read and spontaneous speech. In both cases segmental quality appears to be an important determinant of ratings of overall pronunciation quality. Fluency also appears to be an important aspect that is equally related to all other dimensions investigated.

Second, 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. In particular, in presenting the speech material we suggested that the differences between the items used for the two proficiency groups in Experiment 2 might influence the pronunciation ratings. As explained above, the short and the longs tasks differ not only with respect to length, but also with respect to the nature of the task. More precisely, the BP items contain questions that can be answered immediately by the candidate without much thinking. In general, a given situation is presented and the candidate has to indicate what he/she would say in that context. The IP items, on the other hand, contain questions that require more preparation to be answered. For example, the candidate has to choose between various possibilities and has to explain why he/she made that choice, which means that the candidate, when answering, has to reflect to find good motivations for his/her choice. In other words, the IP items require more cognitive effort than the BP items, which, in turn, could explain the lower pronunciation scores since more cognitively demanding tasks are associated with a lower articulation rate, a lower phonation/time ratio and more pauses (Goldman-Eisler 1968 and Grosjean 1980: 39-53). This is exactly what appears from the comparison of the data for BP and IP in Table 4.

Third, 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 5 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 (Cucchiarini et al. 2000b: 989-999) 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 (Chambers 1997: 535-544) and are corroborated by the data concerning the three proficiency levels: Table 4 shows that the differences between the proficiency levels with respect to *mlp* are relatively smaller than those concerning *#ps*. As already noted in (Cucchiarini et al. 2000b: 989-999) 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 4 shows that the pronunciation ratings are relatively strongly correlated with *ros*, *ptr*, *#ps*, and *mlr*, while *art* and *mlp* have almost no correlation. It is clear that pauses are much more frequent in spontaneous speech than in read speech (see Table 3). 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 observe in Table 5 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. We can also see that the predictive power of *mlr* is greater for SSIP, 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.

5. Conclusions

In this paper we have investigated 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. On the basis of the findings presented and discussed in the previous sections, we can conclude that both in read and spontaneous speech quantitative, temporal measures of speech are strongly related to ratings of pronunciation quality. However, not all variables that appear to be suitable for measuring pronunciation quality in read speech can be employed in spontaneous speech. In particular, variables that measure the rate at which sounds are produced without taking the frequency and the distribution of pauses into account appear to be unsuitable for measuring pronunciation quality in spontaneous speech. Moreover, the importance of the various quantitative measures appears to be dependent on the specific task used to elicit the speech material.

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 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

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

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.

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

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

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

Profieltoets, onderdeel Spreken, June 1998, Arnhem: Cito.

Cucchiarini, C., Strik, H. & Boves, L. (2000b). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology, *Journal of the Acoustical Society of America* Vol. 107 (2), 989-999.

Riggenbach, H. (1991). Toward an understanding of fluency: a microanalysis of nonnative speaker conversations. *Discourse processes* 14: 423-441.

Freed, B.F. (1995). What makes us think that students who study abroad become fluent? In Freed, B.F., (ed.), *Second language acquisition in a study-abroad context.* Amsterdam: John Benjamins, 123-148.

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

Grosjean, F. (1980). 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.

Chambers, F. (1997). What Do We Mean by Fluency? *System*, 4, 535-544