ELSEVIER

Speech Communication 30 (2000) 109–119

SPEECH COMMUNICATION

www.elsevier.nl/locate/specom

Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms

Catia Cucchiarini *, Helmer Strik, Lou Boves

Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

Received 5 February 1998; received in revised form 25 December 1998; accepted 11 February 1999

Abstract

The ultimate aim of the research reported on here is to develop an automatic testing system for Dutch pronunciation. In the experiment described in this paper automatic scores of telephone speech produced by native and non-native speakers of Dutch are compared with specific, i.e., temporal and segmental, and global pronunciation ratings assigned by three groups of experts: three phoneticians and two groups of three speech therapists. The goals of this experiment are to determinutee (1) whether specific expert ratings of pronunciation quality contribute to our understanding of the relation between human pronunciation scores and machine scores of speech quality, (2) whether different expert groups assign essentially different ratings, and (3) to what extent rater pronunciation scores can be predicted on the basis of automatic scores. The results show that collecting specific ratings along with overall ones leads to a better understanding of the relation between human and automatic pronunciation assessment. Furthermore, after normalization no considerable differences are observed between the ratings by the three expert groups. Finally, it appears that the speech quality scores produced by our speech recognizer can predict expert pronunciation ratings with a high degree of accuracy.

Keywords: Automatic pronunciation assessment; Expert ratings; Native and non-native pronunciation

1. Introduction

In the last few years we have witnessed the appearance of numerous software programs for teaching and testing language proficiency, such as those developed by Auralog and Syracuse Language Systems (see URLs in the reference list). The eventual advantages of such systems are obvious: lower costs, greater flexibility and, in the case of testing, increased objectivity.

In developing automatic instruments for language testing it soon appeared that for certain skills automation would be easier than for others. In general four skills are distinguished on the basis of the dimensions: *mode* (oral versus written) and *direction* (receptive versus productive). Since in testing receptive skills it is possible to use response tasks that are easy to score (multiple choice, matching and cloze), developing automatic tests for these skills is relatively easy. For productive skills, on the other hand, automatic tests are difficult to develop, because of the open-ended nature

* Corresponding author. Tel.: +31-24-361-5785; fax: +31-24-361-2907.

E-mail addresses: catia@let.kun.nl (C. Cucchiarini), strik@let.kun.nl (H. Strik), boves@let.kun.nl (L. Boves).

0167-6393/00/$ - see front matter
PII: S0167-6393(99)00040-0

of the input. Furthermore, in the case of speaking, direction and mode conspire to make automatic testing even more difficult.

In spite of these difficulties, various methods for evaluating certain oral sub-skills like pronunciation have been proposed (Bernstein et al., 1990; Neumeyer et al., 1996; Franco et al., 1997). Most of these systems make use of recent developments in automatic speech recognition. However, it seems important that any system intended for testing or improving pronunciation should refer to some standard based on judgments of human raters, the importance of which cannot be overestimated, as human scores are what automatic grading techniques purport to reproduce.

The importance of expert ratings for automatic assessment of pronunciation quality has been underlined by Bernstein et al. (1990). In this study aimed at determinuteing the feasibility of automatic pronunciation grading, the performance of an automatic speech recognizer was tested against speech quality ratings by experts. In (Neumeyer et al., 1996) and (Franco et al., 1997), pronunciation scores assigned by human experts were also used as a reference to determinutee the validity of automatic measures of speech quality such as log-likelihood scores, timinuteg scores, phone classification error scores and segment duration scores. While in these studies considerable effort was dedicated to optimizing the automatic measures so as to obtain better correlations between machine scores and human scores, less attention was paid to the ratings assigned by the experts; only overall ratings of pronunciation were collected.

However, research on pronunciation evaluation has revealed that overall scores of pronunciation quality may be affected by a great variety of speech characteristics (Anderson-Hsieh et al., 1992). Non-native speech can deviate from native speech in various aspects such as fluency, syllable structure, word stress, intonation and segmental quality. When native speakers are asked to score non-native speech on pronunciation quality, their scores are usually affected by more than one of these aspects. Research on the relationship between native speaker ratings of non-native pronunciation and deviance in the various aspects of speech quality has revealed that each area affects the overall score to a different extent (Anderson-Hsieh et al., 1992).

These findings suggest that global ratings of pronunciation quality assigned by human raters have a complex structure, which may be problematic when such scores are used as a reference for automatically produced measures of speech quality, because one does not know exactly what the human scores stand for. Questions such as "What do raters exactly evaluate?" and "What influences their judgements most?" should be taken into consideration when trying to develop machine measures that best approach human pronunciation scores. Against this background it seems that more specific pronunciation ratings should be collected along with global ratings of pronunciation quality so as to obtain a better understanding of pronunciation grading by humans.

Another problem with human pronunciation scores collected in previous studies (Neumeyer et al., 1996; Franco et al., 1997) is that they do not take due account of possible shibboleth sounds. In these studies the experts were asked to assign a global pronunciation score to each of several sentences uttered by each speaker (sentence level rating). The scores for all the sentences by one speaker were then averaged to obtain an overall speaker score (speaker level rating) (see Neumeyer et al., 1996; Franco et al., 1997). Although this procedure may seem logical at first sight, there are some problems with it.

The scores assigned by a rater to different sentences uttered by one and the same speaker may differ as a function of segmental make-up (Labov, 1966). For example, if a shibboleth (stigmatizing) sound is present in one sentence, the score for that sentence may be considerably lower than those for other sentences by the same speaker that do not contain that specific sound. Owing to the presence of a stigmatizing sound, pronunciation score collected at the speaker level could turn out to be lower than the scores that would result by averaging over the various sentences uttered by the same speaker. In other words, the average score might not reflect the effect of the shibboleth sound to the same extent as the one expressed in an overall speaker score. This seems to suggest that if the researcher is interested in pronunciation scores

at the speaker level, (s)he should have the human raters listen to fragments containing the whole phonetic inventory of the language in question.

In our research directed at developing an automatic pronunciation testing system for Dutch, we also took human judgements as a reference. In order to obtain greater insight in how experts evaluate pronunciation, we asked them to assign both global and specific ratings of pronunciation quality. Moreover, in order to take account of the possible effects of stigmatizing sounds on the ratings, in the present experiment the human raters did not assign scores to individual sentences, but judged the pronunciation of each speaker on the basis of two sets of five phonetically rich sentences.

When it came to selecting raters to asses non-native pronunciation of Dutch we found that we could choose from among different groups. Phoneticians are obvious candidates, since they are expert on pronunciation in general. Teachers of Dutch as a second language would seem to be another obvious candidate; however, from these teachers we learned that, in practice, pronunciation problems in learners of Dutch as a second language are not usually addressed by language teachers, but rather by speech therapists. Since it is possible that the ratings vary with the experts' background, we decided to include different groups of raters in the experiment so that we could make comparisons between them.

Another characteristic of the current experiment is that it is not limited to assessing non-native speech, but it also concerns native speech. The reason for doing this is that the presence of native-produced sentences facilitates judgements of non-native speech (Flege and Fletcher, 1992, p. 385). These authors suggest that although native speech patterns are stored in native listeners' long-term memory, the availability of native speech makes it easier for listeners to make accurate judgements of degree of accent.

Finally, an important feature of this experiment is that telephone speech is used. The rationale behind this is that in the future automatic tests to be adminuteistered over the telephone will be required for different applications. In one study that we know of telephone quality was simulated by using 200–3600 Hz band-limited speech (Bernstein et al., 1990). However, this is only a first approximation of real telephone speech.

The first aim of the experiment reported on here was to determinutee whether the availability of specific ratings of pronunciation quality along with global ratings would enhance our understanding of the relation between the human scores and machine scores. The second aim was to determinutee whether resorting to different groups of experts would lead to different results. Finally, we wanted to establish to what extent speech quality scores computed by our speech recognizer (see Strik et al., 1997) can predict pronunciation scores assigned by human experts.

This paper is organized as follows. Section 2 describes the experimental methodology. The results of this experiment are presented and discussed in Section 3, while conclusions are drawn in Section 4.

2. Experimental design

2.1. Speakers

The speakers involved in this experiment are 60 non-native speakers (NNS), 16 native speakers with fairly strong regional accents (NS) and 4 Standard Dutch speakers (SDS). The speakers in the three groups were selected according to different sets of variables. The 60 NNS were selected so as to obtain a group that was sufficiently varied with respect to language background, proficiency level and sex. Similarly, the 16 NS were selected so as to obtain a group of male and female speakers with accent from different regions of the country. Finally, the four speakers of Standard Dutch (two males and two females) were selected on the basis of the high scores they had obtained in previous experiments in which the degree of standardness of their pronunciation had been evaluated (Kraayeveld, 1997).

2.2. Speech material

Each speaker read two sets of phonetically rich sentences (see Appendix A). In preparing the sentences, the following criteria were adopted:

- the sentences should be meaningful and should not sound strange;

- the sentences should not contain unusual words which NNS are unlikely to be familiar with, foreign words or names, or long compound words which are particularly difficult to pronounce;
- the content of the sentences should be as neutral as possible; for instance, the sentences should not contain statements concerning characteristics of particular countries or nationalities;
- each set of five sentences should contain all phonemes of Dutch at least once.

The average duration of each set is about 30 s. With two sets this amounts to roughly 1 minute of speech per speaker. The sentences were read over the telephone. The subjects called from their homes, so that the recording conditions were far from ideal. All speech material was checked and orthographically transcribed before being used for the experiment. Sentences containing disfluencies were transcribed accordingly and were retained in the scoring procedure.

2.3. Expert ratings

A group of three phoneticians (ph) and a group of three speech therapists expert on pronunciation problems of Dutch L2 learners (st1) were selected for this investigation. Moreover, to get greater insight in the degree of reliability that can be attained, a second group of three other speech therapists who are expert on pronunciation problems of Dutch L2 learners (st2) was added.

All raters listened to the speech material and assigned scores individually. No specific instructions were given as to the use of the evaluation scales. 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 help them anchor their ratings. The five speakers were chosen so as to give an indication of the range that the raters could possibly expect.

For each rater the experiment was divided into two parts which were held on different days. In Part 1 the rater assigned overall pronunciation ratings for each set of sentences, while in Part 2 specific ratings were assigned for segmental quality, fluency and speech rate. This setup was chosen to ensure that the overall ratings would not be influenced by the specific ones. Overall pronunciation quality, segmental quality and fluency were rated on a scale ranging from 1 to 10. A scale ranging from −5 to +5 was used to assess speech rate.

For each group of raters, the 80 speakers were proportionally assigned to three raters. Each rater was assigned 20 NNS, 6 NS (2 NS were evaluated twice) and all 4 SDS; 60 sets of sentences (two sets per speaker) were evaluated. Furthermore, in order to be able to compute intrarater and interrater reliability, the raters evaluated 12 sentence sets per scale twice, while 44 sets were evaluated by all three raters in a group.

In the original experimental design the amount of speech material in each part turned out to be too much for two rating sessions. Therefore two tapes were prepared for each part separating the sentence sets (e.g., all speakers' sentences for set 1 were on one tape). The duration of each of the tapes was about 30 minutes. The first tape of each part contained the five training sets that were intended to familiarize the raters with the scoring task. After having rated tape 1, the raters had to pause for a while before starting with tape 2.

In total, 44 sets of sentences of different speakers were scored by all three raters in one group. The scores assigned to these sets were used to calculate interrater reliability for each group. For each rater interrater reliability was calculated on the basis of 12 of these 44 sets that the rater in question had scored twice.

The scores assigned to the two sentence sets produced by each speaker were subsequently averaged to obtain one score per speaker. This way 80 human-assigned scores were obtained for each of the four scales, which were subsequently compared with the various automatic measures.

2.4. Automatic scores

In this experiment the speech recognizer described in (Strik et al., 1997) was used. As the recording system was connected to an ISDN line, the recorded signals consist of 8 kHz 8 bit A-law coded samples. Feature extraction is done every 10 ms for frames with a width of 16 ms. The first step in feature analysis is a Fast Fourier Transform to calculate the spectrum. In the following step, the

energy in 14 mel-scaled filter bands between 350 and 3400 Hz is calculated. Next, a discrete cosine transformation is applied to the log filter band coefficients. The final processing stage is a running cepstral mean substraction. Besides 14 cepstral coefficients ($c_0 - c_{13}$), 14 delta coefficients are also used. This makes a total of 28 feature coefficients.

The continuous speech recognizer (CSR) uses acoustic models (39 Hidden Markov Models, HMMs), language models (unigram and bigram), and a lexicon. The lexicon contains orthographic and phonemic transcriptions of the words to be recognized. The HMM consist of three segments of two identical states, one of which can be skipped. One HMM was trained for non-speech sounds and one for silences. For each of the phonemes /l/ and /r/ two models were trained, a distinction was made between prevocalic (/l/ and /r/) and postvocalic position (/L/ and /R/). For each of the other 33 phonemes one HMM was trained.

The CSR was trained by using part of the Polyphone corpus (den Os et al., 1995). This corpus is recorded over the telephone and consists of read and (semi-)spontaneous speech of 5000 subjects with varying regional accents. For each speaker 50 items are available. Five of these 50 items are phonetically rich read sentences, which contain all phonemes of Dutch at least once. Each speaker read a different set of sentences. In this experiment the phonetically rich sentences from 4019 speakers were used for training the CSR.

By applying a form of forced Viterbi alignment we computed two temporal measures of speech quality, which had proven to be successful (Cucchiarini et al., 1997): TD is the total duration of speech plus pauses, ROS the rate of speech (total number of segments/TD).

For the present research we added another automatic measure, namely a likelihood ratio (LR). Likelihood ratios have been employed in previous research for utterance verification, i.e., to test whether a certain utterance has been spoken (see, e.g., Lee, 1997). It is therefore worthwhile to investigate whether likelihood ratios can be used to determinutee whether a subject realized the prompted utterance with a poor pronunciation. The ratios used by Lee (1997) 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, independently of speech rate measures. The LR used in the present research is simply the difference of two log-likelihood scores obtained with different versions of the CSR:

$LR = LL1 - LL2$

LL1: Forced Viterbi alignment with the canonical transcriptions and the 39 HMMs (37 monophone HMMs, 1 noise and 1 silence HMM)

LL2: Free phone recognition with the same 39 HMMs, using the phone language models during decoding (i.e., applying loose phonotactic constraints). The phone language models (unigram and bigram) were trained on the Polyphone material.

The general idea behind these log-likelihood measures is the following: for good pronunciation LL1 and LL2 are very similar and their ratio (LR) is small; for poor pronunciation LL1 is smaller; but not LL2, so that the ratio of the two (LR) is negative and its absolute value is higher than that for good pronunciation.

For all three automatic measures we calculated one average score per speaker by averaging over the scores for the two sentence sets of each speaker. This way three sets of 80 scores were obtained, which were then compared with the expert ratings.

3. Results

The data collected in the present experiment were analyzed in order to answer the three questions posed in the introduction: (1) whether specific ratings of pronunciation quality contribute to a better understanding of expert pronunciation scoring, (2) whether different experts produce different ratings, and (3) whether expert pronunciation ratings can be predicted on the basis of automatically obtained scores. However, before studying the relationship between global and specific ratings of pronunciation it is essential to know whether ratings are at all reliable and whether they vary with the rater group. We therefore start this

section by presenting the results concerning the reliability of the expert ratings and the comparisons between the rater groups (3.1). In Section 3.2 we focus on the relationship between the different rating types. Finally, Section 3.3 deals with the relation between the expert ratings and the automatic scores.

3.1. Expert ratings: reliability and comparison between expert groups

On the basis of the sets of sentences that each rater evaluated twice (24 scores), intrarater reliability could be established for the four scales overall pronunciation (OP), segmental quality (SQ), fluency (FL) and speech rate (SR). The results for the three raters in each group are shown in Table 1.

Although there are some differences between the various scales and the various raters, in general intrarater reliability appears to be satisfactory.

For each group of experts interrater reliability was calculated on the basis of the 44 sets of sentences that were evaluated by all three raters in each group (132 scores). 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, even in the least favourable condition (C3), the reliability coefficients are still rather high. Subsequently, we checked the degree of correlation between the ratings assigned by the three rater groups. The results are shown in Table 3.

Since agreement errors are known to affect the size of the correlation coefficient, the correction for attenuation formula was applied (Ferguson, 1987), 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. This is not surprising if we consider that the raters did not receive instructions on the use of the scales.

The agreement within a group of raters has obvious consequences for the correlation coefficient computed between the combined scores

Table 1
Intrarater reliability (Cronbach's α) for the various scales (OP, SQ, FL and SR) and the raters in the three groups

	Phoneticians			Speech therapists 1			Speech therapists 2		
	R1	R2	R3	R4	R5	R6	R7	R8	R9
OP	0.97	0.95	0.99	0.85	0.94	0.97	0.93	0.92	0.98
SQ	0.96	0.98	0.93	0.86	0.98	0.99	0.74	0.94	0.95
FL	0.97	0.94	0.95	0.94	0.97	0.96	0.90	0.76	0.91
SR	0.94	0.76	0.74	0.73	0.84	0.88	0.85	0.94	0.72

Table 2
Interrater reliability (Cronbach's α) for three rater groups in three different conditions: C1, C2, and C3

	Phoneticians			Speech therapists 1			Speech therapists 2		
	C1	C2	C3	C1	C2	C3	C1	C2	C3
OP	0.97	0.96	0.89	0.95	0.93	0.89	0.95	0.93	0.87
SQ	0.97	0.97	0.92	0.95	0.93	0.85	0.90	0.84	0.74
FL	0.96	0.95	0.96	0.93	0.91	0.88	0.90	0.88	0.83
SR	0.86	0.84	0.87	0.82	0.76	0.81	0.84	0.82	0.84

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

	OP	SQ	FL	SR
ph–st1	0.92	0.90	0.94	0.90
ph–st2	0.80	0.57	0.82	0.88
st1–st2	0.90	0.69	0.83	0.81

of the raters and another set of data (i.e., the ratings by another group or the machine scores). This is so, because straightforward combination of the scores would amount to pooling measurements made with different yardsticks. When such an inhomogeneous set of measurements is submitted to a correlation analysis with homogeneous measures, the 'jumps' at the splicing joints lower the correlation. The same is true 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.

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 (44 sentence sets), because in this case all raters scored the same samples. For individual raters, these values hardly differed from the means and standard deviations for the total material.

The effect of normalizing the data is evident from Table 4 which shows the correlations between the three groups of raters after normalization. These correlations are higher than those in Table 3. Moreover, they are so high that we can conclude that all nine raters involved in this experiment judge pronunciation in a similar way. Given the advantages of normalization, standard scores will be used also in the rest of the analyses in this study.

The results presented in this section show that expert raters manage to evaluate global and specific aspects of pronunciation quality with a high degree of reliability, even though they did not receive instructions on how to use the various scales. Agreement, on the other hand, was not very high, probably due to the lack of instructions. Although the degree of agreement does play a crucial role in constructing a pronunciation test, because it contributes to establishing the cutoff point, in the present experiment reliability was our main concern, while agreement was less important, because we are still in the development stage.

Table 4
Correlations between the ratings of the three rater groups (ph, st1, st2) after normalization

	OP	SQ	FL	SR
ph–st1	0.96	0.91	0.94	0.93
ph–st2	0.90	0.87	0.90	0.86
st1–st2	0.94	0.84	0.90	0.89

One of the aims of this experiment was to determinutee whether asking different groups of experts to judge pronunciation quality would lead to essentially different ratings. On the basis of the results presented above we can conclude that, after normalization of the data, no considerable differences can be observed between the ratings by the different experts. This is an important finding and, in a way, also a reassuring one, because it indicates that expert pronunciation scores do have a certain degree of stability.

3.2. Overall and specific ratings: results and discussion

By comparing the overall pronunciation scores with the specific ones, it is possible to establish how the separate aspects of pronunciation quality investigated here are related to overall pronunciation. In general, all correlations in Table 5 are relatively high, although there are slight differences. The highest correlations are found between overall pronunciation and segmental quality, which suggests that when the raters judge overall pronunciation, they are most influenced by the quality of the segments uttered by the speaker. This appears to be in line with the general idea that the term pronunciation refers to the degree of correctness in articulating individual sounds.

The finding that these correlations are high is amenable to two different interpretations: either the various aspects of pronunciation quality are indeed highly correlated with each other, in which case the raters used the scales correctly, or the raters failed to score the various aspects independently of each other. The second interpretation seems less plausible if we consider that overall pronunciation and segmental quality were evaluated in two

Table 5
Correlations between the four scales for the three rater groups

		SQ	FL	SR
OP	ph	0.97	0.87	0.73
	st1	0.96	0.87	0.60
	st2	0.91	0.77	0.64
SQ	ph		0.86	0.69
	st1		0.91	0.61
	st2		0.76	0.62
FL	ph			0.87
	st1			0.83
	st2			0.83

separate sessions and still their correlation is very high. It also appears that a clearly temporal scale such as speech rate is more strongly correlated with fluency, which should also be related to temporal characteristics of speech, than with segmental quality. However, the fluency scale seems to occupy an intermediate position, since it is almost equally correlated with the three other scales. On the whole these results seem to suggest that the raters did their job properly: when asked to rate fluency and speech rate, they indeed paid attention to aspects of speech timinuteg. In other words, the high correlations between the four types of human-assigned scores are most probably due to the fact that these aspects of pronunciation quality are indeed intertwined. A comparison of the human ratings with the machine scores may shed more light on this point.

3.3. Automatic correlates: results and discussion

In calculating the correlations between the various automatic measures no correction for attenuation was applied, because in this case $\alpha = 1.00$ (in repeating the calculations exactly the same scores would be obtained). The correlation between TD and ROS is very high (−0.96), while both variables are moderately correlated with LR (−0.65 and 0.63, respectively). This was to be expected, because TD and ROS are temporal measures, while LR should be more related to the spectral characteristics of speech.

We now go on to consider the correlations (corrected for attenuation) between the four types of human ratings of the three expert groups and the three automatic measures. The results presented in Table 6 show that the two temporal measures TD and ROS are more strongly correlated with the human ratings than LR. Among the human-assigned scores, fluency shows the highest correlations with the automatic scores for all expert groups, while ROS appears to be the best predictor of the human ratings.

Table 6
Correlations between the automatic measures and the standard scores by the three rater groups (ph, st1, st2)

		OP	SQ	FL	SR
TD	ph	−0.79	−0.75	−0.91	−0.90
	st1	−0.81	−0.77	−0.94	−0.88
	st2	−0.73	−0.70	−0.91	−0.88
ROS	ph	0.82	0.79	0.93	0.92
	st1	0.83	0.79	0.91	0.89
	st2	0.77	0.76	0.90	0.89
LR	ph	0.49	0.45	0.64	0.62
	st1	0.54	0.55	0.68	0.65
	st2	0.47	0.44	0.55	0.59

Another thing to be noted is that specific aspects of pronunciation quality can be predicted more accurately than overall pronunciation, provided that the right automatic correlate is found. In other words, although overall pronunciation can be predicted accurately on the basis of ROS and TD, it appears that ROS can predict fluency and speech rate even more accurately, which is also what one would expect. These results also corroborate our impression that the raters used the scales in the proper way, because their ratings of timinuteg are indeed more strongly correlated with objective temporal measures than their ratings of overall pronunciation and segmental quality.

Segmental quality turns out to be the variable that is predicted most poorly on the basis of the automatic scores, which is not a positive finding if we consider that segmental quality is the best predictor of overall pronunciation. A remarkable result in Table 6 is that LR is more correlated with fluency and speech rate than with overall pronunciation and segmental quality, while it was intended to be a correlate of the spectral characteristics of speech. This suggests that to achieve better automatic scoring of pronunciation quality we should try to find a more adequate correlate of segmental quality.

In order to test the contribution of the various automatic measures, we performed stepwise multiple regression analyses in which OP was the criterion and the automatic scores were the predictors. No increase in the multiple correlation coefficient ($R = 0.82$) was observed when LR was entered in the regression equation after ROS. In trying to interpret these results we should not forget that the correlation between ROS and OP is already very high. In any case these findings suggest that LR does not contain information that is not present in the ROS scores. This could be due to the fact that speech rate and pronunciation quality are very closely related in our data, which is borne out by the correlations in Table 5. 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.

It is indeed conceivable 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 intertwined.

An important question that remains to be answered is why the measure LR did not show a stronger correlation with SQ and OP. First of all, it should be noted that in this experiment we worked with two broad classes: there was a lot of variation both in the speakers L1 background and in the speech material used to train the CSR. Since our instrument has to measure how well foreign speakers from a wide range of L1 backgrounds speak Dutch, we did not limit the experiment to learners with one specific L1, but included subjects with various mother tongues. In addition, we did not want to limit the reference to Standard Dutch, but wanted to allow all generally accepted regional Dutch accents. The result of this large amount of variation on both sides is that in our experiment the task is fairly complex, more complex than, for instance, in the situation described in (Franco et al., 1997), where the task was to judge speech from a relatively homogeneous group (Native Americans) who have to speak Parisian French. A second explanation, which is partly related to the first one, is that our LR is not optimal. As mentioned above, LRs are best computed by comparing the LL of a model with that of a specific anti-model: LR = LL (model)/LL (anti-model) (Lee, 1997). In our situation it is not obvious a priori what the optimal anti-model should be. Should we use L1 speech for many languages to train the anti-model or should we use non-native Dutch produced by speakers with different L1? Even if we knew how the anti-model should be trained, we would not have the necessary database at our disposal. Given that we could not define an optimal anti-model, we decided to use a different approach, i.e., we calculated log-likelihoods with two different versions of the CSR.

At this point it is interesting to compare our results with those of previous studies. The most obvious comparison is that with the research conducted by Neumeyer et al. (1996) and by Franco et al. (1997). However, before comparing their results with ours, a few things should be pointed out.

First of all, these authors collected ratings for each individual sentence of each speaker. The scores of all sentences of one speaker were subsequently averaged. This way two sorts of scores were obtained: (1) sentence-level scores and (2) speaker-level scores. Correlations at sentence level are consistently lower than those at speaker level. Since we do not have scores at sentence level, the first type of score cannot be used for comparisons. As to the speaker-level scores, Neumeyer et al. (1996) average over 30 sentence scores per speaker and Franco et al. (1997) over 50 sentence scores. In our case, the scores at speaker level are obtained by averaging over two scores per speaker, each of which is based on five sentences by the same speaker.

Second, our correlation coefficients are corrected for attenuation. Since we do not have the reliability coefficients of their expert ratings to correct their correlation coefficients, we will compare their correlation coefficients with ours prior to correction.

In (Neumeyer et al., 1996) the best predictor of expert ratings appeared to be segment duration

scores: $r = 0.85$. In (Franco et al., 1997) posterior probabilities turned out to have marginally greater predictive power: $r = 0.88$. The highest correlation we found between the automatic measures and overall pronunciation was 0.80 (st1, without correction) while for the fluency ratings the highest coefficient was −0.93 for TD. It seems therefore that these coefficients are of the same order of magnitude.

Furthermore, it is worth mentioning that our findings warrant the use of overall ratings of pronunciation as sole reference for the automatic scores. As a matter of fact, the specific ratings collected in this study appeared to be highly correlated with overall ratings.

Another aspect in which our study differs from previous ones is that telephone speech was used. People were asked to dial a certain number and they were free to select time, place and location for placing the call. Consequently, the resulting acoustic registrations differ in many ways from those made in a studio or a (usually quiet) office environment. Here we will mention only the most relevant ones.

First of all, in telephone speech only the bandwidth of 300–3400 Hz is used. Second, not just one high quality microphone was used, but many different telephone microphones. Finally and probably most important, relatively high level acoustic background signals are frequently present, which is usually not the case with laboratory speech. We do consider these conditions as 'normal and realistic', in the sense that later on, when this technology will be used in applications over the telephone, conditions will most probably be similar. However, it should be underlined that these conditions make automatic speech recognition more difficult.

4. Conclusions

The first aim of the experiment reported on in this paper was to find out whether specific ratings of pronunciation quality would increase our insight into the relation between human ratings and machine scores. The results presented above show that this is indeed the case: the comparison between more detailed and global ratings revealed that overall pronunciation is most influenced by segmental quality, which is the human measure that can be predicted most poorly on the basis of our machine scores. It also appeared that specific aspects of pronunciation quality can be predicted more accurately, provided that the right automatic correlate is found. In other words, although overall pronunciation can be predicted accurately on the basis of automatic measures of timinuteg, it appears that these measures can predict fluency and speech rate even more accurately, which is also what one would expect. A clear result of this experiment is that the optimal correlate of segmental quality still eludes us.

It seems therefore that an important contribution of the specific ratings is that they make clear in which direction action should be taken in order to achieve better pronunciation scoring. For example, it is now clear that attempts should be made to obtain a better predictor of segmental quality, because this would prevent speakers with poor pronunciation and the right temporal characteristics from obtaining high pronunciation scores.

The second aim of this experiment was to determinutee whether taking different groups of experts as a reference would lead to different ratings. The results presented above reveal that raters who did not receive any instructions on the use of the rating scales may differ from each other in the absolute values of the scores assigned. However, one can normalize for these differences by computing standard scores. After normalization no considerable differences between the raters were observed: they all evaluate the speakers in a similar way. We can therefore conclude that expert ratings of pronunciation exhibit a certain degree of stability.

Finally, the third aim of the experiment reported on in this paper was to determinutee whether pronunciation ratings assigned by human experts can be predicted on the basis of scores produced by an automatic speech recognizer. The results found so far show that a good prediction of both global and specific pronunciation scores can be obtained on the basis of automatic measures of timinuteg such as ROS and TD. However, it seems that further research is needed to determinutee whether appropriate measures can be found to obtain a more refined assessment of segmental quality.

With a view to the ultimate aim of our research, i.e., developing an automatic testing system for

Dutch pronunciation, the results of this experiment are very useful since they show that pronunciation scores assigned by human experts can be accurately predicted on the basis of measures computed by a speech recognizer. Furthermore, they indicate how we should proceed toward developing an automatic pronunciation test. For instance, finding an adequate automatic correlate of segmental quality is necessary to avoid that fast speaker with low proficiency get high pronunciation scores.

To conclude, the results presented in this paper are promising and the fact they were obtained under rather 'normal and realistic' conditions (no laboratory speech, no exclusion of disfluent utterances) makes them even more promising.

Acknowledgements

This research was supported by SENTER (which is an agency of the Dutch Minuteistry of Economic Affairs) under the Information Technology Programme, the Dutch National Institute for Educational Measurement (CITO), Swets Tests 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. We thank Febe de Wet for her assistance in analyzing the data.

Appendix A

Set 1

1. Vitrage is heel ouderwets en past niet bij een modern interieur.
2. De Nederlandse gulden is al lang even hard als de Duitse mark.
3. Een bekertje warme chocolademelk moet je wel lusten.
4. Door jouw gezeur zijn we nu al meer dan een uur te laat voor die afspraak.
5. Met een flinke garage erbij moet je genoeg opbergruimte hebben.

Set 2

1. Een foutje van de stuurman heeft het schip doen kapseizen.
2. Gelokt door een stukje kaas liep het muisje keurig in de val.
3. Het ziet er naar uit dat het deze week bij ons opnieuw gaat regenen.
4. Na die grote lekkage was het dure behang aan vervanging toe.
5. Geduldig hou ik de deur voor je open.

References

Anderson-Hsieh, J., Johnson, R., Koehler, K., 1992. The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. Language Learning 42, 529–555.

Auralog, http://www.auralog.com/eng/index.htm.

Bernstein, J., Cohen, M., Murveit, H., Rtischev, D., Weintraub, M., 1990. Automatic evaluation and training in English pronunciation. In: Proceedings International Congress on Spoken Language Processing (ICSLP) '90, Kobe, pp. 1185–1188.

Cucchiarini, C., Strik, H., Boves, L., 1997. Automatic evaluation of Dutch pronunciation by using speech recognition technology. In: Furui, S., Juang, B.H., Chou, W. (Eds.), Proceedings IEEE Workshop, ASRU, Santa Barbara, pp. 622–629.

den Os, E.A., Boogaart, T.I., Boves, L., Klabbers, E. 1995. The Dutch polyphone corpus. In: Pardo, J.M., Enríquez, E., Ortega, J., Ferreiros, J., Macías, J., Valverde, F.J. (Eds.), Proceedings ESCA Fourth European Conference on Speech Communication and Technology: EUROSPEECH 95, Madrid, pp. 825–828.

Ferguson, G.A., 1987. Statistical Analysis in Psychology and Education, fifth edition. McGraw-Hill, Singapore.

Flege, J., Fletcher, K., 1992. Talker and listener effects of perceived foreign accent. J. Acoust. Soc. Amer. 91, 370–389.

Franco, H., Neumeyer, L., Kim, Y., Ronen, O., 1997. Automatic pronunciation scoring for language instruction. In: Werner, B. (Ed.), Proceedings International Congress on Acoustics, Speech and Signal Processing (ICASSP) 1997, München, pp. 1471–1474.

Kraayeveld, H., 1997. Idiosyncrasy in prosody. Doctoral dissertation University of Nijmegen, Nijmegen.

Labov, W., 1966. The social stratification of English in New York City. Center for Applied Linguistics, Washington.

Lee, C.H., 1997. A unified statistical hypothesis testing approach to speaker verification and verbal information verification. In: Proceedings COST Workshop, Rhodos, pp. 63–72.

Neumeyer, L., Franco, H., Weintraub, M., Price, P. 1996. Automatic text-independent pronunciation scoring of foreign language student speech. In: Bunnel, H.T., Idsardi, W. (Eds.), Proceedings International Congress on Spoken Language Processing (ICSLP) '96, Philadelphia, pp. 1457–1460.

Strik, H., Russel, A., Heuvel, H., Cucchiarini, C., Boves, L., 1997. A spoken dialogue system for the Dutch public transport information service. Internat. J. Speech Technol., 121–131.

Syracuse Language Systems, http://www.syrlang.com/.