AUTOMATIC ASSESSMENT OF SECOND LANGUAGE LEARNERS' FLUENCY

Catia Cucchiarini and Helmer Strik
A^{2}RT, Dept. of Language and Speech, University of Nijmegen, the Netherlands
{cucchiarini, strik}@let.kun.nl, http://lands.let.kun.nl/

ABSTRACT

This paper describes an experiment aimed at determining whether automatic assessment of second language learners' fluency in spontaneous speech is feasible and whether it differs from automatic fluency assessment in read speech. Spontaneous speech of 60 learners of Dutch was scored for fluency by five raters and was analyzed by means of a continuous speech recognizer to calculate seven quantitative measures of speech quality known to be related to perceived fluency. The results show that automatic assessment of second language learners' fluency in spontaneous speech is feasible, although not all variables suitable for measuring fluency in read speech are as effective in spontaneous speech. In particular, measures that express the rate at which sounds are produced without taking pauses into account appear to be unsuitable for measuring fluency in spontaneous speech. Furthermore, the correlations between machine scores and human ratings are lower for spontaneous speech. Possible explanations are discussed.

1. INTRODUCTION

Oral fluency is viewed as an important characteristic of second language speech, which is often the object of evaluation in testing second language skill. For this reason attempts have been made to try to define fluency in terms of objective properties of speech, since this would contribute to more objective fluency testing [1, 2, 3]. Some of these attempts have made use of a dual approach in which perceived fluency scores assigned by raters to non-native speech are compared with objective measures calculated for the same speech [1, 2, 3]. These studies have revealed that perceived fluency is particularly affected by factors such as speech rate and pauses, while self-repairs are a poor fluency indicator.

In a previous paper [4] we reported on an experiment in which such a dual approach was adopted. This study differed from previous ones in two important respects: First, the objective measures that were calculated manually in previous studies [1, 2, 3], were calculated automatically by means of a continuous speech recognizer (CSR); second, instead of using spontaneous speech we decided to use read speech, so that the raters would not be distracted by differences in grammar and vocabulary, which are known to affect fluency ratings [2].

In the study reported on in [4] read speech of 20 native and 60 non-native speakers of Dutch was scored for fluency by nine experts and was then analyzed by means of a CSR in terms of quantitative fluency measures such as speech rate, articulation rate, number and length of pauses, number of dysfluencies, mean length of runs and phonation-time ratio. As explained above, the decision to limit this investigation to read speech was related to the methodological complexities involved in studying fluency in spontaneous speech . However, the idea was to apply this approach to spontaneous speech too, if it turned out to be feasible for read speech.

This experiment produced interesting results in two respects, a. fluency assessment by expert raters and b. the relationship between expert fluency ratings and automatically obtained objective fluency measures. With regard to a., the results showed that expert ratings of fluency in read speech are reliable (Cronbachs' α varies between .90 and .96). With respect to b., very high correlations were found between the expert fluency ratings and the automatic measures of fluency: five automatic measures showed correlations with the fluency scores whose magnitude varied between .77 and .91. The highest correlations were found for rate of speech (between .86 and .91). Further analyses revealed that two factors are important for perceived fluency in read speech: the rate at which speakers articulate the sounds and the number of pauses they make. Rate of speech appears to be such a good predictor of perceived fluency because it incorporates these two aspects.

Since automatic assessment of fluency in read speech turned out to be feasible, we decided to extend this approach to spontaneous speech. To pursue this aim, we used a test developed by the Dutch National Institute for Educational Measurement (Cito), the *Profieltoets* [5], which Cito administered in June 1998. This test contains items which elicit unprepared answers that can therefore be classified as extemporaneous, spontaneous speech. As in the experiment in [4], the speech material was evaluated by a group of raters and by the automatic speech recognizer. The methodology and the results of this experiment are described in the rest of this paper.

The aim of the present paper is to determine whether automatic assessment of language learners' fluency in spontaneous speech is feasible and whether and in which respects it differs from automatic fluency assessment in read speech.

2. METHOD

2.1. Speakers

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 (LP) at the level of 'basic user' and a higher proficiency group (HP) at the level of 'independent user' The 30 speakers in each group were randomly selected from a larger group of candidates who hadparticipated in the test. Within each of these two groups the speakers varied with respect to sex and mother tongue.

2.2. Speech Material

The speech material used in this experiment consists of the answers given by the above-mentioned candidates to a subset of eight items stemming from the *Profieltoets*. These eight items were specifically selected because they elicit unprepared answers, so that the speech can be characterized as extemporaneous. Moreover, given the nature of the questions, reasonably long answers can be expected, a necessary condition for calculating fluency measures. In the items selected for the LP group the subjects were allowed to talk for 15 s, whereas the HP subjects had 30 s at their disposal. Effectively, the LP subjects talked for about 70 s on average, while for the HP subjects the average was 170 s.

In order to be analyzed by the CSR, the speech material was orthographically transcribed. Repetitions, restarts, repairs and filled pauses were also transcribed.

2.3. Fluency Ratings

In the study reported on in [4] phoneticians and speech therapists functioned as raters, phoneticians because they are experts on pronunciation in general and speech therapists because their expertise is usually invoked when learners of Dutch exhibit pronunciation problems (including all fluency-related temporal phenomena). In the present experiment ten teachers of Dutch as a second language were employed because they are normally used as raters for this kind of examination by Cito.

All raters listened to the speech material and assigned scores individually. They could listen to the speech fragments as often as they wanted. For eight items they were asked to score fluency on a scale ranging from 1 to 10. As in the experiment in [4], no specific instructions were given for fluency assessment. For each speaker involved in this experiment we therefore obtained five fluency scores assigned by five raters.

An essential difference between the two experiments is that in the present one two different groups of raters were assigned to the two groups of speakers, whereas in [4] the same group of raters evaluated all speakers. As a consequence, in [4] we could compare different groups of speakers on the perceived fluency scores, whereas this is not possible in the present experiment.

2.4. Automatic Assessment of Fluency

In this experiment the automatic speech recognizer described in [7] was used. 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 (FFT) to calculate the spectrum. The energy in 14 mel-scaled filter bands between 350 and 3400 Hz is then calculated. Next, a discrete cosine transformation is applied to the log filterband coefficients. The final processing stage is a running cepstral mean subtraction. Besides 14 cepstral coefficients (c0 - c13), 14 delta coefficients are also used. This makes for a total of 28 feature coefficients.

The CSR uses acoustic models (39 Hidden Markov Models) and a lexicon. The lexicon contains orthographic and phonetic transcriptions of the words to be recognized. The continuous density HMMs 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 distincion was made between prevocalic and postvocalic position. For each of the other 33 Dutch phones one model was trained. The HMMs were trained by using the phonetically rich sentences of 4019 speakers from the Polyphone corpus [8].

For each individual answer the automatic measures were calculated by means of a form of forced Viterbi alignment. The number of phones was determined on the basis of the transcriptions. The various fluency scores for the individual items were subsequently averaged over the eight items. This way a set of 57 (the data of three subjects turned out to be missing) scores was obtained for each measure, which were then compared with the human fluency scores.

By means of the CSR a number of quantitative measures known to be related to perceived fluency were calculated. On the basis of the results from the literature on the use of temporal variables in studying speech production [1, 2, 3, 5], the following measures were selected for investigation:

- *ros* = rate of speech: # segments / total duration of speech plus sentence-internal pauses
- *ptr* = phonation/time ratio: total duration of speech without pauses / total duration of speech plus sentence-internal pauses
- *art* = articulation rate : # segments / total duration of speech without pauses
- *tdp* = total duration of sentence-internal pauses: all silences longer than or equal to 0.2 s
- *alp* = average length of pauses
- *#p* = # of silent pauses
- *mlr* = mean length of runs: average number of phones between unfilled pauses of not less than 0.2 s.

3. RESULTS

In this section the results of the present experiment are presented. In section 3.1. we report the results concerning the fluency ratings assigned by the two groups of raters. In 3.2. we examine the results concerning the quantitative measures of fluency. Finally, in 3.3. the correlations between these two types of results are considered.

3.1. Fluency Ratings

The fluency scores assigned by the two rater groups RLP (raters for the LP group) and RHP (raters for the HP group) were analyzed to determine interrater reliability (see Table 1).

	interrater reliability
RLP	.86
RHP	.82

Table 1. Interrater reliability coefficients (Cronbach's α) for the two rater groups, RLP and RHP.

As is clear from Table 1, interrater reliability is reasonably high. On the one hand, this may be surprising if we consider that the raters involved in this experiment were given no specific instructions for assessing fluency. On the other, these reliability coefficients are lower than those in the previous experiment [4], but this can be explained. First, here we analyze the two groups of speakers separately, with the consequence that the variance is much lower. Second, in [4] we had deliberately chosen read speech material so that the raters would not be distracted by differences in grammar and vocabulary, which are known to affect fluency ratings [2]. For instance, if we compare these results with those of previous studies in which spontaneous speech was employed [2, 3], then we may conclude that our reliability cofficients are relatively high.

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 five raters in each group. Therefore, we decided to normalize for the differences in the values by using standard scores instead of raw scores, as was done in [4].

3.2. Quantitative Measures of Fluency

In this section we analyze the quantitative variables in various respects. First of all, we calculate the mean and standard deviation for all variables for all groups, because this may be helpful in interpreting the various correlations later on. To get a better understanding of the behavior of the different variables, we also compare these means and standard deviations with those of the read speech data for natives and non-natives. Table 2 contains these results.

	read speech				spontaneous speech			
	natives		non-natives		LP		HP	
	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
ros	12.74	1.35	9.68	1.94	5.99	0.96	5.31	1.17
ptr	93.17	2.79	82.66	8.57	49.32	8.71	44.92	9.51
art	13.65	1.19	11.61	1.37	12.25	1.25	11.85	.81
#p	1.42	1.23	7.20	5.47	36.28	11.14	94.52	22.84
tdp	0.45	0.42	3.10	2.76	32.41	9.26	93.51	27.23
alp	0.20	0.13	0.38	0.13	0.92	0.19	1.02	0.28
mlr	34.26	5.85	21.52	8.77	9.50	2.22	9.33	2.27

Table 2 Means and standard deviations for the seven quantitative measures for read speech of natives and non-natives and spontaneous speech of LP and HP.

Table 2 shows how the values for the different variables vary as a function of speech modality (read vs. spontaneous) and speaker group. Of course we should bear in mind that the quantity of speech material varied in the various conditions, with the result that relative measures such as *ros*, *art*, *ptr*, *alp* and *mlr* are comparable, whereas absolute measures such as *#p* and *tdp* are not.

In order to see how the temporal measures vary as a function of speech modality we can compare the read speech values for the non-natives (columns 4 and 5) with the spontaneous speech values of the two (also non-native) speaker groups (columns 6, 7, 8 and 9).

The most striking differences between read and spontaneous speech concern the variables *ros*, *ptr*, *alp* and *mlr*. In particular, if we go from read speech to spontaneous speech we observe that the values for *ros* and *ptr* are almost halved, while that of *alp* isalmost tripled. *art*, on the other hand, hardly changes. However, *art* does vary if we go from natives to non-natives in the read speech group.

In order to allow comparisons for the two measures *#p* and *tdp*, we normalized them for total duration of speech plus sentence-internal pauses, thus obtaining two new variables *#pn* and *tdpn*. *tdpn* is actually redundant, because it is the complement of *ptr*, and will not be presented here. For *#pn* we found the following values for the four groups:

		$\bar{x}$	sd
read speech	natives	0.09	0.08
	non-natives	0.31	0.24
spontaneous speech	LP	0.52	0.16
	HP	0.53	0.13

Table 3. Means and standard deviations for #pn for read speech of natives and non-natives and spontaneous speech of LP and HP.

The data presented above (Tables 2 and 3) 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.

3.3. Fluency Ratings and Quantitative Measures

In this section we compare the fluency scores assigned by the raters with the automatically calculated temporal measures of speech, in order to determine to what extent the latter are able to predict the former. To this end the degree of correlation between the two sets of scores was calculated. Since the ratings assigned to the groups of speakers are not comparable (because they were assigned by different raters), the correlations will be calculated for each group separately. In this way the variation in proficiency level is reduced, with obvious consequences for the correlations.

In calculating these correlations the correction for attenuation formula was applied, because this makes it possibe to correct for measurement errors, which are known to affect the size of the correlation coefficient. Moreover, this allows comparisons between the various coefficients. In order to be able to make comparisons with the read speech data, we also present the correlations between the quantitative measures computed for the read speech material and the fluency ratings of the three rater groups: phoneticians (Ph), speech therapists 1 (St1) and speech therapists 2 (St2). These results are shown in Table 4.

Table 4 reveals that the correlations between the fluency ratings for spontaneous speech and the automatic fluency measures are very different for the various measures. For instance, *ros*, *ptr* and *mlr* exhibit relatively high correlations with the ratings of both groups, whereas *art* and *alp* shows no relation with the automatic measures. The correlations for *#p* and *tdp* are considerably different for the two groups.

The low correlations between art and the fluency ratings may be attributed to the low variance in the *art* scores, which can be inferred from the data presented in Table 2. However, the same argument cannot be used to explain the low correlations between alp and the fluency ratings, because the variance in *alp* is comparable to that in *#p* and *tdp*. So, the absence of a relation between *alp* and the fluency ratings is still very surprising.

At this point it may be interesting to compare these correlations with those obtained for read speech in the experiment described in [4]. These data are presented in Table 4 in columns 2, 3 and 4.

	read speech			spontaneous speech	
	Ph	St1	St2	RLP	RHP
ros	.93	.91	.90	.61	.43
ptr	.86	.88	.89	.49	.43
art	.88	.84	.81	.07	.05
#p	-.84	-89	-.89	-.56	-.32
tdp	-.81	-.86	-.86	-.68	-.25
alp	-.65	-.62	-.65	-.09	-.01
mlr	.85	.86	.88	.53	.72

Table 4. Correlations (corrected for attenuation) between the fluency ratings and the quantitative measures for spontanoeus speech and for read speech, for different rater groups.

A comparison between the two data sets reveals that also in the read speech data *alp* showed the lowest correlation with the fluency ratings, so in a sense there is some correspondence between the two sets of data on this point. However, it remains to be explained why for spontaneous speech *alp* shows no relation at all with the fluency ratings.

Furthermore, the comparison between read speech and spontaneous speech reveals that the correlations are much higher in the first case. This is not surprising if we consider that in the read speech experiment there was much more variation in proficiency level than in the present experiment. First, in the previous experiment native speakers were also included, while the spontaneous speech data were limited to non-natives. Second, the lower amount of variation in the present experiment is further reduced because we have to analyze the two groups separately. So, if we consider the substantial differences with respect to the amount of variation, we have to conclude that the correlations observed in the present experiment are not bad at all, at least for some of the quantitative measures.

In addition, the enormous differences between the read speech material and the spontaneous speech material with respect to the quality of the recordings should be kept in mind. First of all, the spontaneous speech material was recorded under rather adverse environmental conditions: 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. Second, many subjects spoke so softly that in certain cases it was almost impossible to understand what they said. Furthermore, they produced repetitions, restarts and repairs. In a sense it is a pity that dysfluencies and filled pauses were not calculated for this material. One can imagine that although these variables were no good predictors of fluency in read speech, they might play an important role in spontaneous speech, simply because they are more common. It is our intention to carry out these analyses in the near future.

To summarize, if we consider all the factors mentioned above, then one may wonder how the CSR managed to segment this speech material at all. Concurrent speech, background noise, etc. are known to degrade the performance of CSRs to a great extent. So, the lower correlations may simply be due to the considerably larger amount of ‘noise’ in these data compared to the read speech data.

In spite of all these adverse factors some automatic measures, in particular *ros, mlr* and *ptr*, appear to be rather stable indicators of fluency. What these measures have in common is that they are all complex variables that express some kind of relation between speech and silence, and it is probably this relation that is at the core of perceived fluency.

4. CONCLUSIONS

In this paper we have investigated the feasibility of automatic assessment of second language learners’ fluency in spontaneous speech and have compared these results with those obtained in a previous experiment for read speech. On the basis of the findings presented and discussed in the previous sections, we can conclude that automatic assessment of second language learners’ fluency in spontaneous speech is feasible, although not all variables that appear to be suitable for measuring fluency 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 length of pauses into account appear to be unsuitable for measuring fluency in spontaneous speech.

ACKNOWLEDGMENTS

This research was supported by the Ministry of Economic Affairs (SENTER), the Dutch National Institute for Educational Measurement (Cito), Swets and Zeitlinger and PTT Telecom. The research of Dr. H. Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. We would like to thank Uriël Schuurs and Ron Engelen of Cito, Arnhem, for making the scores of the Profieltoets available to us, and for their advice in the data processing.

REFERENCES

[1] Lennon, P. 1990. Investigating fluency in EFL: a quantitative approach. *Language Learning* 40: 387-417.

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

[3] 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.

[4] Cucchiarini, C., Strik, H. & Boves, L. 1998. Quantitative assessment of second language learners' fluency: an automatic approach, *Proceedings ICSLP98*, 30 nov. - 4 dec., Sydney, Australia, Vol. 6, 2619-2622.

[5] Profieltoets, onderdeel Spreken, June 1998, Arnhem: Cito.

[6] Grosjean, F., 1980. Temporal variables within and between languages. In Dechert, H.W. & M. Raupach, (eds.), *Towards a cross-linguistic assessment of speech production*, Lang, Frankfurt, 39-53.

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

[8] den Os, E.A., Boogaart, T.I., Boves, L., and Klabbers, E. 1995. The Dutch Polyphone corpus, *Proc. Eurospeech95*, Madrid: 825-828.