QUANTITATIVE ASSESSMENT OF SECOND LANGUAGE LEARNERS' FLUENCY: AN AUTOMATIC APPROACH

Catia Cucchiarini, Helmer Strik and Lou Boves

A²RT, Dept. of Language and Speech, University of Nijmegen, the Netherlands

ABSTRACT

This paper describes an experiment aimed at determining whether native and non-native speakers of Dutch significantly differ on a number of quantitative measures related to fluency and whether these measures can be successfully employed to predict fluency scores. 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 an automatic speech recognizer in order to calculate nine quantitative measures of speech quality that are known to be related to perceived fluency. The results show that the natives' scores on the fluency ratings and on the quantitative measures significantly differ from those of the non-natives, with the native speakers being considered more fluent. Furthermore, it appears that quantitative variables such as rate of speech, phonation-time ratio, number of pauses, and mean length of runs are able to predict fluency scores with a high degree of accuracy.

1. INTRODUCTION

The term fluency is commonly used by second language teachers and researchers to describe speech production performance of second language learners. This suggests that there is general agreement as to the precise meaning of fluency. However, a review of relevant literature reveals that this term has been used to refer to different skills in different contexts [1, 2, 3, 4].

In an attempt to gain more insight into this concept, studies were carried out [1, 2, 3] in which speech samples were scored for fluency by experts and were then analyzed in terms of several temporal variables. These studies reveal that perceived fluency is particularly affected by factors such as speech rate and pauses, while self-repairs are a poor fluency indicator. Moreover, the findings suggest that quantitative analysis may be useful in distinguishing between more and less fluent speech and in determining fluency improvements. In turn this would suggest that this type of research may contribute to developing objective fluency testing instruments and, possibly, automatic fluency tests.

However, it must be pointed out that the results of the studies mentioned above only indicate trends that should be verified with larger samples of speakers, as the authors themselves suggest [1, 2, 4], because in these investigations small numbers of speakers (4 in [1], 6, in [2] and 8 in [3]) were involved. Furthermore, these studies had some other shortcomings. For instance, since spontaneous speech was used, the speech samples could vary along many dimensions (grammar, pronunciation, vocabulary etc.). These factors are known to affect fluency ratings [2]. This might in part explain the low degree of reliability observed between the raters [2, 3]. Moreover, only non-native speakers were involved, while comparison with native speakers would be necessary to establish norm ranges that are required for testing purposes [1].

The present research aims at gaining more insight into the factors that affect perceived fluency, while at the same time addressing some of the shortcomings of previous studies. In this investigation 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 an automatic speech recognizer in order to calculate quantitative measures of speech quality that are known to be related to perceived fluency. By using this dual approach we hope to arrive at a clearer definition of what constitutes fluency in read speech.

Another aim of the present study is to find out whether natives and non-natives significantly differ on fluency ratings and on a number of quantitative variables related to perceived fluency. Finally, we want to determine whether quantitative variables can be successfully used to predict fluency scores of read speech.

2. METHOD

2.1. Subjects and Speech Material

The speakers involved in this experiment are 60 non-native speakers (NNS) and 20 native speakers of Dutch (NS). 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 20 NS were selected in order to obtain a heterogenous group with respect to region of origin and sex.

Each speaker read two sets of five phonetically rich sentences. The average duration of each set is 30 s. With two sets this amounts to one minute of speech per speaker. All speech material was orthographically transcribed before being used for the experiment.

The sentences were read over the telephone. As the recording system was connected to an ISDN line, the input signals consist of 8 kHz 8 bit A-law coded samples. The subjects called from their homes or from telephone booths, so that the recording conditions were far from ideal. Since one of the aims of this experiment was to determine whether fluency can be automatically scored, because this would be advantageous for testing, we decided to use telephone speech so that we could also determine whether this type of testing would be possible through the telephone.

2.2. Expert Fluency Ratings

For the aim of assessing non-native fluency different experts could be used as raters. Phoneticians are obvious candidates, because they are experts on pronunciation in general. Teachers of Dutch as

a second language would seem to be another obvious choice. However, it turned out that, in practice, delivery problems of learners of Dutch are usually addressed by specially trained speech therapists, who, therefore, would seem to better qualify as 'non-native speech experts' than language teachers. Finally, three groups of raters were selected. The first group consisted of three expert phoneticians (ph) with considerable experience in judging pronunciation and other speech and speaker characteristics. The second and the third groups each consisted of three speech therapists (st1 and st2) who had considerable experience in treating students of Dutch with pronunciation problems.

All raters listened to the speech material and assigned scores individually. They could listen to the speech fragments as often as they wanted. Fluency was rated on a scale ranging from 1 to 10. No specific instructions were given for fluency assessment. However, five sets of sentences spoken by five different speakers were played to the raters before they started with the evaluation proper, so as to help them anchor their ratings.

In order to limit the amount of material to be scored by each rater, the 80 speakers were proportionally assigned to the three raters in each group. The scores assigned by the three raters were then combined to compute correlations with the automatic scores and between rater groups. In order to compute intrarater and interrater reliability, 12 sentence sets by different speakers were evaluated twice by each rater while 44 sentence sets were scored by all three raters in each group.

2.3. Automatic Assessment of Fluency

In this experiment the automatic speech recognizer described in [6] was used. This ASR was trained by using the phonetically rich sentences of the Polyphone corpus [7]. By means of the ASR 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, 8, 9], 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 sec
- alp = average length of pauses
- #p = # of silent pauses
- mlr = mean length of runs: average number of phones occurring between unfilled pauses of not less than 0.20 secs
- #fp = # filled pauses: ə, əm
- #dy = # dysfluencies (repetitions, restarts, repairs)

3. RESULTS

In this section the results of the present experiment are presented in the following order. In section 3.1. we report the results concerning the fluency ratings assigned by the three groups of experts. In 3.2. we look at the results concerning the quantitative measures of fluency. Finally, in 3.3 the correlations between the se two types of results are considered.

3.1. Expert Fluency Ratings

The fluency scores assigned by the three rater groups were analyzed to determine intrarater and interrater reliability (see Table 1).

	intrarater reliability			**interrater reliability**
	rater 1	**rater 2**	**rater 3**	
ph	.97	.94	.95	.96
st1	.94	.97	.96	.93
st2	.90	.76	.91	.90

Table 1 Intrarater and interrater reliability coefficients (Cronbach's alpha) for the three rater groups, ph, st1, and st2.

As is clear from Table 1, both intrarater and interrater reliability are very high. Only for rater 2 of the second group of speech therapists is intrarater reliability considerably lower than for all other raters, but it is still within acceptable limits. These results clearly differ from those of previous studies, in which lower degrees of reliability were reported, probably because raters adopted different definitions of fluency [2, 3].

Besides considering interrater reliability, we also checked the degree of interrater agreement. Closer inspection of the data revealed that the means and standard deviations varied between the raters in a group, but also between the raters in different groups who rated the same speech material. The agreement within a group of raters has obvious consequences for the correlation coefficient computed between the combined scores of the raters and another set of data (i.e. the ratings by another group or the quantitative variables). 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 scores), because in this case all raters scored the same samples. Within the individual raters the values for the 44 overlapping samples hardly differed from the means and standard deviations for the total material. Table 2 shows the correlation coefficients between the groups of raters before and after normalization. It is known that measurement errors affect the size of the correlation coefficient; therefore, the correction for attenuation formula was applied, so as to allow comparisons between the various coefficients.

	Raw scores	Standard scores
ph - st1	.92	.94
ph - st2	.82	.90
st1 - st2	.83	.90

Table 2 Correlations between the rater groups before and after normalization.

From Table 2 it appears that normalization has the effect of enhancing the degree of correlation between the groups, as was to be expected. Given the advantages of normalization, standard scores will be used in the rest of the analyses in this study.

In order to determine whether natives and non-natives significantly differ on the expert fluency ratings, the standard scores of the three rater groups were submitted to a *t*-test for equality of means. The results of this test are shown in Table 3.

	$\bar{x}$ ns	sd ns	$\bar{x}$ nns	sd nns	*t*-value	df	p
ph	.88	.39	-.32	.70	9.55	59.98	.000
st1	.91	.13	-.27	.79	11.07	67.55	.000
st2	.86	.33	-.30	.83	8.90	75.77	.000

Table 3 Results of *t*-test for the fluency ratings of the three rater groups.

As appears from Table 3, the mean scores assigned to the two speaker groups are very similar for the three rater groups. Furthermore, the two groups of NS and NNS significantly differ on the ratings assigned by the three rater groups, with the native speakers being considered more fluent than the non-natives. It is clear that not only the mean scores differ considerably between the two speaker groups, but also the standard deviations, thus indicating that the group of NS is more homogeneous in this respect than the group of NNS.

3.2. Quantitative Measures of Fluency

Similarly, the quantitative measures of fluency were analyzed to determine whether significant differences could be observed between the two groups of natives and non-natives. Table 4 shows that the two groups do indeed differ significantly on all measures. These results may contribute to the discussion on the usefulness of temporal variables in distinguishing between natives and non-natives. Although it is true that native speech is not always perfectly smooth and continuous [2], it appears that, on average, native speech exhibits fewer pauses and dysfluencies, while speed of delivery is higher than in non-native speech. Moreover, these results are in line with those of previous studies that investigated the speech performance of the same speakers in both L1 and L2 and that were based on smaller samples [5, 9].

Table 4 reveals that the number of filled pauses and dysfluencies is extremely low. This is not surprising if we consider that we are dealing with read speech and that these phenomena are known to occur rarely in oral reading [9]. This suggests that these features may be no good indicators of fluency in read speech.

	$\bar{x}$ ns	sd ns	$\bar{x}$ nns	sd nns	*t*-value	df	p
ros	12.74	1.35	9.68	1.94	6.54	78	.000
ptr	93.17	2.79	82.66	8.57	11.07	67.55	.000
art	13.65	1.19	11.61	1.37	5.97	78	.000
#p	1.42	1.23	7.20	5.47	-7.62	73	.000
tdp	0.45	0.42	3.10	2.76	-7.18	66.68	.000
alp	0.20	0.13	0.38	0.13	-5.236	78	.000
mlr	34.26	5.85	21.52	8.77	7.359	49.20	.000
#fp	0.00	0.00	0.14	0.35	-3.18	59	.002
#dy	0.12	0.22	0.62	0.76	-4.49	77.4	.000

Table 4 Results of *t*-tests for the nine quantitative measures.

3.3. Fluency Ratings and Quantitative Measures

In the preceding sections we have shown that natives and non-natives differ significantly both on fluency ratings and on a set of quantitative variables that are supposed to be related to perceived fluency. However, these results are not sufficient to conclude that the machine-derived variables are indeed good fluency indicators. To find out whether this is the case, the degree of correlation between the fluency ratings and the quantitative variables has to be calculated. The results of these analyses are shown in Table 5.

	Phoneticians	Speech therapists 1	Speech therapists 2
ros	.93	.91	.90
ptr	.86	.88	.89
art	.88	.84	.81
#p	-.84	-89	-.89
tdp	-.81	-.86	-.86
alp	-.65	-.62	-.65
mlr	.85	.86	.88
#fp	.34	.33	.38
#dy	.42	.48	.40

Table 5 Correlations (corrected for attenuation) between the fluency ratings by the three rater groups and the quantitative measures.

From Table 5 it appears that all tempo-related variables are strongly correlated with fluency ratings, with the exception of alp. On the other hand, hesitation phenomena such as filled pauses and dysfluencies show no strong correlation with fluency scores. This latter finding is probably related to the fact that these phenomena are so rare in the type of speech under investigation (see Table 4).

4. DISCUSSION AND CONCLUSIONS

In this paper we have presented the results of a study on fluency in which a dual approach was adopted: fluency ratings assigned by experts to read speech produced by natives and non-natives were compared with a number of temporal measures calculated for the same speech fragments. The innovations of this study are the following. First, more subjects were involved than in previous investigations. Second, automatic speech recognition technology was used to compute the quantitative measures, which has important advantages concerning the objectivity of the measurements and the amount of data that can be handled. Third, both native and non-native speakers were involved. Fourth, the use of read speech made it possible to rule out the influence of some linguistic factors known to affect fluency ratings [2], while concentrating exclusively on purely acoustic variables.

The results show that reliability was high for all three groups of experts (Cronbachs' α varied between .90 and .96), while high agreement was obtained by using standard scores. 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 and that in previous studies low degrees of reliability were obtained [2, 3]. On the other, 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]. So, the contrast with previous studies might be due to a difference in the type of speech material being evaluated: read speech vs. spontaneous, conversational speech.

Native and non-native speakers appear to differ significantly on the fluency ratings and on all quantitative measures. These findings are interesting in the light of the discussion on the effectiveness of temporal variables in distinguishing between native and non-native speakers. Although it is true that not all native speakers are completely fluent [2], these results show that, on average, they are more fluent, produce fewer pauses and dysfluencies, and speak faster than non-native speakers. In turn this suggests that the quantitative variables employed in this experiment may be successfully used to distinguish between natives and non-natives.

With respect to the definition of fluency, these results show that, at least for read sentences, speed of delivery, as expressed by measures such as rate of speech, articulation rate, phonation time ratio and mean length of runs, is a very good fluency indicator.

The results presented above also show that automatic scoring of fluency in read speech is possible: as the correlations between five of the tempo-related measures and the expert ratings vary between -.81 and .93, it can be concluded that fluency scores can be predicted with a high degree of accuracy. This conclusion seems to be even more warranted if we consider that the correlations between the fluency ratings of the experts varied between .90 and .94. In other words, the correlations between the expert ratings and the automatic fluency scores are very similar to those between the ratings of different expert groups.

To conclude, these findings suggest that the use of temporal measures of speech production together with automatic speech recognition techniques may contribute to developing automatic tests of fluency, at least for read speech. If we then consider that these results were obtained with telephone speech, it may be legitimate to conclude that this approach has enormous potentials for the future of fluency assessment.

5. 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.

6. REFERENCES

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

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

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

4. Chambers, F. "What do we mean by fluency?", *System*, Vol. 25, No. 4: 535-544, 1997.

5. Towell, R., Hawkins, R., and Bazergui, N., "The development of fluency in advanced learners of French", *Applied Linguistics*, 17: 84-119, 1996.

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

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

8. Grosjean, F., "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, 1980.

9. Möhle, D., "A comparison of the second language speech production of different native speakers", in: Dechert, H.W., Möhle, D., & Raupach, M., (eds.) *Second language productions*, Narr, Tübingen, 26-49, 1984.