

# Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology

Catia Cucchiarini, Helmer Strik, and Lou Boves

*A<sup>2</sup>RT, Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500HD Nijmegen, The Netherlands*

(Received 15 December 1998; revised 13 July 1999; accepted 29 September 1999)

To determine whether expert fluency ratings of read speech can be predicted on the basis of automatically calculated temporal measures of speech quality, an experiment was conducted with read speech of 20 native and 60 non-native speakers of Dutch. The speech material was scored for fluency by nine experts and was then analyzed by means of an automatic speech recognizer in terms of quantitative measures such as speech rate, articulation rate, number and length of pauses, number of dysfluencies, mean length of runs, and phonation/time ratio. The results show that expert ratings of fluency in read speech are reliable (Cronbach's  $\alpha$  varies between 0.90 and 0.96) and that these ratings can be predicted on the basis of quantitative measures: for six automatic measures the magnitude of the correlations with the fluency scores varies between 0.81 and 0.93. Rate of speech appears to be the best predictor: correlations vary between 0.90 and 0.93. Two other important determinants of reading fluency are the rate at which speakers articulate the sounds and the number of pauses they make. Apparently, rate of speech is such a good predictor of perceived fluency because it incorporates these two aspects. © 2000 Acoustical Society of America.

[S0001-4966(00)04401-5]

PACS numbers: 43.70.Kv, 43.71.Es, 43.71.Gv, 43.71.Hw [JMH]

## INTRODUCTION

The term fluency is routinely used by teachers and researchers to describe both native and non-native language performance. The fact that fluency is a frequently applied notion might suggest that there is general agreement as to its precise meaning. However, a review of relevant literature reveals that the term fluency has been used to refer to a wide range of different skills and different speech characteristics (e.g., Leeson, 1975; Fillmore, 1979; Brumfit, 1984; Lennon, 1990; Schmidt, 1992; Chambers, 1997).

In spite of this great variation, though, there is general agreement on two matters. First, although it is obvious that fluency can be used to describe written performance (Lennon, 1990), most authors restrict the use of the term to the oral modality. Furthermore, although some authors have underlined the importance of fluency-related factors in receptive processes (Leeson, 1975; Segalowitz, 1991), there seems to be a tacit agreement among teachers and researchers that fluency mainly refers to productive language performance. However, even this more restricted definition of fluency as a descriptor of oral production is amenable to different interpretations.

In considering the various possibilities, we may draw a distinction between fluency with respect to native language performance and fluency in the context of foreign language teaching and testing. In the latter case, fluency is viewed as an important criterion by which non-native performance can be judged (Riggenbach, 1991), despite the vagueness of the exact meaning of the concept. This is clear from the fact that fluency is often included in tests and evaluation schemes. With respect to native speakers' oral performance, fluency may be used to characterize the performance of a speaker, but does not really constitute an evaluation criterion. The

term dysfluent, on the other hand, is often used in connection with certain speech disorders such as stuttering, where dysfluent speech is characterized by "an abnormally high frequency and/or duration of stoppages in the forward flow of speech" (Peters and Guitar, 1991).

In considering native speakers' oral production Fillmore (1979) identifies four different abilities that might be subsumed under the term fluency: (a) "the ability to talk at length with few pauses," (b) "the ability to talk in coherent, reasoned, and "semantically dense" sentences," (c) "the ability to have appropriate things to say in a wide range of contexts," and (d) "the ability...to be creative and imaginative in...language use."

In foreign language teaching and testing, various definitions of fluency are also found. For instance, in communicative language teaching the emphasis has been on fluency as opposed to accuracy. According to the definition provided by Brumfit (1984, p. 57) fluency is "the maximally effective operation of the language system so far acquired by the student." In this definition of fluency, native-speaker-like performance does not constitute the target to be achieved (Brumfit, 1984, p. 56). Alternatively, nativelylike performance is viewed as the final goal in the more common interpretation of fluency as a synonym for oral command of a language. In everyday language use, this definition may be extended to indicate overall language proficiency (Lennon, 1990; Chambers, 1997). Finally, in a more restricted sense, the term fluency has been used to refer to one aspect of oral proficiency, in particular the temporal aspect (Nation, 1989; Lennon, 1990; Riggenbach, 1991; Schmidt, 1992; Freed, 1995; Towel *et al.*, 1996). However, even when the term fluency is used in this more limited sense, there is still uncertainty as to what exactly contributes to perceived fluency. It is this—

admittedly rather vague—temporal interpretation of fluency that will be the focus of the present paper.

In trying to define the temporal aspect of fluency, it has often been assumed that the goal in language learning consists of producing “speech at the tempo of native speakers, unimpeded by silent pauses and hesitations, filled pauses...self-corrections, repetitions false starts and the like” (Lennon, 1990). However, quantitative studies of pause-related phenomena have revealed that native speech is not always smooth and continuous, but exhibits a lot of hesitations and repairs (Raupach, 1983; Lennon, 1990; Riggenbach, 1991). This would seem to imply that the presence of hesitation phenomena is not sufficient to distinguish between natives and non-natives and that the difference rather lies in the frequency and distribution of these phenomena, as suggested by Möhle (1984). As a matter of fact, studies that have compared a number of quantitative fluency measures in L1 and L2 speech of the same speaker have shown that there may be considerable differences between the two speech types (Möhle, 1984; Towell *et al.*, 1996).

In an attempt to gain more insight into the temporal aspects of fluency, Lennon (1990), Riggenbach (1991), and Freed (1995) carried out studies in which samples of spontaneous speech produced by non-native speakers of English were judged by experts on fluency and were then analyzed in terms of quantitative variables such as speech rate, phonation-time ratio, mean length of runs, and number and length of pauses. The results of these studies show that fluency ratings are affected by quantitative variables such as speech rate and number of pauses. In addition, these studies also reveal that studying the relationship between fluency ratings and temporal variables in spontaneous speech may be rather complex, because in this case the fluency ratings turn out to be affected by nontemporal properties of speech utterances, such as grammar, vocabulary, and accent (Lennon, 1990; Riggenbach, 1991; Freed, 1995).

The aim of the research reported in this paper is to determine whether expert fluency ratings of read speech can be predicted on the basis of temporal measures of speech quality. The decision to limit this investigation to read speech is related to the methodological complexities involved in studying fluency in spontaneous speech. If the present approach appears to be feasible, it will be applied to spontaneous speech too. Identifying quantitative correlates of perceived fluency is important with a view to developing objective testing instruments for fluency assessment. An important characteristic of the present investigation is that the quantitative variables are calculated automatically. In turn this suggests that if the objective measures used in this study appear to be able to predict perceived fluency, this approach may have potential for the development of automatic tests of fluency in read speech.

The goal of this study will be pursued by relating expert fluency ratings of speech read by native and non-native speakers of Dutch with a set of quantitative measures of speech quality that are supposed to be related to perceived fluency. In this way it can be determined to what extent expert judgments of fluency can be predicted on the basis of automatically obtained temporal measures of speech quality.

In other words, the expert fluency ratings will constitute the reference for the evaluation of the automatic fluency measures. Of course, this will be possible only if the expert ratings exhibit acceptable levels of reliability. To this end, we will ask different groups of raters to evaluate the same material on fluency. Moreover, each rater will be asked to score part of the material twice so that it will be possible to establish reliability.

In addition, these analyses will make it possible to determine the contribution of the various quantitative variables to perceived fluency. In turn this will shed some light on the determinants of fluency in read speech.

Furthermore, since the data gathered in this investigation concern both natives and non-natives, this will offer the possibility of determining whether native and non-native speakers differ on the fluency ratings and on the temporal variables. It is clear that distinguishing between these two groups is not the aim of a fluency test, which, instead, should distinguish between fluent and nonfluent speakers. However, for the development of a test of this kind, data on native performance are necessary to establish benchmarks. Moreover, given that fluency is often equated with nativelike performance (see above), it is interesting to determine whether the two groups of natives and non-natives significantly differ from each other on the variables under study.

## I. METHOD

### A. Speakers

The speakers involved in this experiment are 60 non-native speakers (NNS) and 20 native speakers of Dutch (NS). The 60 NNS 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.

Table I shows how the 60 non-native speakers were distributed according to these three variables. Some comment about this table is in order. First, the speakers in the “beginner” category had been attending the course for some months. This was thought to be necessary for the learners to be able to read the sentences. Second, it is clear from this table that the speakers were not evenly distributed over the categories. This has to do with the availability of the speakers. Even if it were possible to find the same number of speakers for each category, then they have to be prepared and have to find the time to carry out the task. So, eventually, there were more women and more speakers of the intermediate and advanced levels. Furthermore, the number of speakers differed for the various mother tongue groups. It is clear, though, that for the purpose of the present experiment, complete symmetry in the sample is not really required.

Four of the NS subjects, two men and two women, were speakers of the Standard variety of Dutch (SDS: Standard Dutch Speakers), while the other 16 NS, speaking an accented variety of Dutch, were selected to obtain a heterogeneous group with respect to region of origin and gender. The rationale behind including the four SDS is that the presence of clear “anchor stimuli” has been shown to be an important help in keeping the reference standard stable (Flege and

TABLE I. Distribution of the 60 non-native speakers according to the selection variables mother tongue, proficiency level, and gender.

	Beginner		Intermediate		Advanced		Total
	Female	Male	Female	Male	Female	Male	
Arabic	1		1	2	1	1	6
Turkish	1			1	1		3
Chinese/Japanese		1	1	1			3
Spanish/Italian/Portuguese	1		3	1	5	2	12
Russian/Polish/Serbo Cr.	1		3	1	5	2	12
English	2		1	1		2	6
German	1		3	2	2	1	9
French		1	2	1	1		5
Swedish/Danish/Norwegian	1		3				4
Total	8	2	17	10	15	8	60

Fletcher, 1994). However, we do not expect the SDS and the NS to be different with respect to fluency, so in the analyses they will be treated as one group of native speakers.

### B. Speech material

Each speaker read two different sets of five phonetically rich sentences designated group 1 and group 2 (see Appendix). In preparing the sentences, the following criteria were adopted:

- (i) the sentences should be meaningful and should not sound strange;
- (ii) 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;
- (iii) 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;
- (iv) each set of five sentences should contain all phonemes of Dutch at least once.

The average duration of each set is 30 s. With two sets this amounts to 1 min of speech per speaker. All speakers read the same sentences over the telephone. The sentences to be read were printed on paper together with the instructions. Consequently, the subjects had the possibility of rehearsing before reading the sentences over the telephone. They had not explicitly been encouraged to do so, but since they had received the material beforehand, they had the chance to rehearse. Moreover, they had the possibility of starting the recording session all over again if they felt something had gone wrong. However, this happened only in one case.

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 were allowed to call from their homes, from telephone booths, or from the first author's office. Two subjects resorted to the latter possibility, while all the others called from their homes. Since the recordings did not take place in sound-treated booths, the recording conditions were different from those in a studio.

All speech material was checked and orthographically transcribed before being used for the experiment. Although

with read speech the content of the sentences should be known beforehand, one cannot be sure that the speaker will read exactly what is on paper. Furthermore, speakers may repeat part of the words or sentences, and make restarts and repairs.

In transcribing the material, special symbols were used for four categories of nonspeech acoustic events,

- (a) filled pauses: uh, er, mm, etc.
- (b) speaker noise: lip smack, throat clear, tongue click, etc.
- (c) intermittent noise: noise that occurs incidentally during the call such as door slam and paper rustle.
- (d) stationary noise: continuous background noise that has a rather stable amplitude spectrum such as road noise or channel noise.

Repetitions, restarts, and repairs were transcribed exactly as they were pronounced. The transcriptions were carried out at SPEX (SPEX), a university expertise center that specializes in database construction and validation.

### C. Raters

Since in this experiment a specific aspect of speech production had to be evaluated, raters with a high level of expertise were required. Different categories of raters seemed to qualify as experts: phoneticians, because they are expert on pronunciation in general; teachers of Dutch as a second language (L2) for obvious reasons. However, it turned out that, in practice, pronunciation problems (including all fluency-related temporal phenomena) of people learning Dutch as L2 are usually not addressed by language teachers, but by specially trained speech therapists. Since it is possible that the ratings vary with the background of experts, a group of three phoneticians and a group of three speech therapists, expert on pronunciation problems of Dutch L2 learners, were selected for this investigation.

Furthermore, since previous studies had revealed that the reliability of expert fluency ratings was rather low (Lennon, 1990; Riggensbach, 1991; Freed, 1995), we decided to add a third group of experts to get more information on the degree of reliability that can be attained. It turned out that finding speech therapists for this task was easier than finding phoneticians, so the third group of experts consisted of three

TABLE II. Distribution of the speech material among the three raters in each group. The cells in italics contain the material used for determining inter-rater reliability, while the material in bold was used for intrarater reliability calculation.

		Material for comparison man-machine				Added material for reliability analyses						
		Total 1				Total 2			Duplications for intrarater reliability		Grand total sum 1 and 2	
Rater 1	group 1	20 NNS1	6 NS1	4 <i>SDS</i>	30	<i>13 NNSA1</i>	<i>5 NSA1</i>	18	<b>5NNSD11</b>	<b>1NSD11</b>	48	
	group 2	20 NNS1	6 NS1	4 <i>SDS</i>	30	<i>14 NNSA2</i>	<i>4 NSA2</i>	18	<b>4NNSD12</b>	<b>2NSD12</b>	48	
Rater 2	group 1	20 NNS2	6 NS2	4 <i>SDS</i>	30	<i>13 NNSA1</i>	<i>5 NSA1</i>	18	<b>5NNSD21</b>	<b>1NSD21</b>	48	
	group 2	20 NNS2	6 NS2	4 <i>SDS</i>	30	<i>14 NNSA2</i>	<i>4 NSA2</i>	18	<b>4NNSD22</b>	<b>2NSD22</b>	48	
Rater 3	group 1	20 NNS3	6 NS3	4 <i>SDC</i>	30	<i>13 NNSA1</i>	<i>5 NSA1</i>	18	<b>5NNSD22</b>	<b>1NSD31</b>	48	
	group 2	20 NNS3	6 NS3	4 <i>SDC</i>	30	<i>14 NNSA2</i>	<i>4 NSA2</i>	18	<b>4NNSD32</b>	<b>2NSD32</b>	48	
<i>132(44×3) sets for inter-rater reliability analyses</i>											<b>36(12×3) sets for inter-rater reliability</b>	

other speech therapists who are expert on pronunciation problems of Dutch L2 learners.

#### D. Expert fluency ratings

The speech material was transferred from disc on a DAT tape adopting different orders for the different raters, as will be described below. All raters listened to the speech material and evaluated it individually. This was done to enhance flexibility (each rater could thus carry out the task at the most suitable time) and to avoid raters influencing each other.

Each rater received two tapes which contained the group 1 and the group 2 sentences, respectively. The material was scored on a scale ranging from 1 to 10. 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 assess fluency. 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 80 speakers were proportionally assigned to the three raters in each group. Each rater was assigned 20 NNS, 6 NS with regional accents (since there were only 16 of these speakers, 2 of them were scored by two raters instead of by only one) and all 4 speakers of the standard variety. For each speaker, two sets of sentences (group 1 and group 2) had to be evaluated, which makes 60 sets of five sentences for each listener. Furthermore, 36 sentence sets were added to allow calculation of intrarater reliability and inter-rater reliability.

In assigning speakers to raters, we took the selection variables into account to avoid overloading raters with speakers of one gender, L1, or level of proficiency. The way in which the speakers were divided over the various raters is illustrated in Table II. Each rater scored the same 20 NNS, the same 6 NS and all 4 SDS twice, once for the group 1 sentences and once for the group 2 sentences, so that 30 scores per rater per sentence group were obtained. The speakers were presented in different random orders in the

two sentence groups, to minimize possible ordering effects on the scores. However, the four SDS were presented at regular intervals, so that the raters would be reminded of how the sentence was supposed to sound in the standard language, as was explained above (see also Flege and Fletcher, 1994). In Table II the distribution of the speakers is clarified by distinguishing three groups of 20 NNS (one for each rater) i.e., 20 NNS1, 20 NNS2, 20 NNS3, and three groups of 6 NS, 6 NS1, 6 NS2, and 6 NS3. Since the four SDS were scored by all three raters in a group, both for the group 1 and the group 2 sentences, the same label 4 SDS is used in Table II for all three raters. The scores assigned by the raters to this part of the material were subsequently compared with the automatic measures calculated for the same material. For this reason this material will be referred to as the man-machine comparison material.

The 36 sentence sets that were added for calculating inter-rater and intrarater reliability were selected so as to have a balanced set of NNS and NS and of group 1 and group 2 sentences. The sentence sets produced by the four SDS were also included in the inter-rater reliability analyses, because they had been scored by all three raters in a group. Consequently, we did not need to add extra SDS sentence sets. Eventually, we selected 27 NNS sets and 9 NS sets and 18 group 1 sets and 18 group 2 sets, as is clear from Table II, under added material. The 13 NNS and the 5 NS sets selected for group 1 and the 14 NNS and the 4 NS sets selected for group 2 were the same for all raters, so the labels 13 NNSA(dded) 1(group 1), 5 NSA1, 14 NNSA2, and 4 NSA2 are used in Table II for all raters.

The number of sentence sets that were eventually used for inter-rater reliability analyses amounts to 44 (36 extra plus the 4 SDS for group 1 and the 4 SDS for group 2, indicated in italic in Table II) per rater, i.e., 132 for all three raters, as appears from the italic cell in the bottom row of Table II.

For the intrarater reliability analyses, on the other hand, 12 sentence sets that were present both in the man-machine comparison material and in the inter-rater reliability material were chosen for each rater. The 12 sets to be scored twice by each rater were selected so as to have nine NNS and three NS and six group 1 sets and six group 2 sets, as appears from the bold cells in Table II, under duplicated materials. Given

that the five NNS and one NS in group 1 and the four NNS and the two NS in group 2 differed for the three raters, different labels are used, i.e., 5 NNSD (uplicated) 1(rater 1)1(group 1), 1 NSD11, 4 NNSD12, 2 NSD12, 5 NNSD21, 1 NSD21, 4 NNSD22, 2 NSD22, 5 NNSD31, 1 NSD31, 4 NNSD32, 2 NSD32.

To summarize, each rater had to evaluate 30+18 sets of sentences (the 6 sets for intrarater reliability were a subset of these 48 sets) of group 1 and 30+18 sets of sentences of group 2. These numbers are indicated in the Total 1 and Total 2 columns in Table II, as well as the grand total for each rater for each group, 48. Since this amount of material was too much for one rating session, it was divided over two sessions. Therefore, two tapes were prepared, one containing 48 sets of sentences of group 1 and the other containing 48 sets of sentences of group 2. The duration of each of the tapes was about 30 min. The first tape contained the five training sets mentioned above. After having rated tape 1, the raters had to pause for a while before starting with tape 2.

The scores assigned to the two sets of sentences by each speaker were subsequently averaged to obtain one score for each speaker. The scores assigned by the three raters were then combined to compute correlations with the machine scores. This way 80 human-assigned fluency scores were obtained, which were subsequently compared with the various quantitative measures.

## E. Automatic assessment of fluency

### 1. The automatic speech recognizer

To calculate the quantitative measures, the continuous speech recognizer (CSR) described in Strik *et al.* (1997) 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 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 continuous density HMMs consist of three parts of two identical states, one of which can be skipped. One HMM was trained for nonspeech sounds and one for silence. 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 HMMs were 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 the so-called phonetically rich 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 of 4019 speakers were used for training the CSR.

The trained CSR was subsequently used to analyze the utterances read by the 80 speakers. For each utterance a Viterbi alignment between the speech signal and the orthographic transcription was obtained. This Viterbi alignment is also a segmentation at the phone level and contains information about the boundaries of phones. Consequently, the segmentation contains information about the position of speech and nonspeech parts (pauses, dysfluencies, etc.). The accuracy of forced alignment was checked only for a small sample of the material. In general the segmentation appeared to be correct, although the boundaries were not always placed where a human listener would probably have placed them. This aspect, however, is not really crucial for the present article, because here we do not use the information about the position of the phone boundaries in the speech parts, but we are concerned with the automatic calculation of the phonemes present in an utterance. This calculation was determined on the basis of the transcriptions, i.e., it is the number of units actually produced and not the number of units the speakers were supposed to realize on the basis of the text they had to read. The resulting segmentation was used to calculate a number of quantitative measures that are described in detail below.

### 2. Quantitative measures of fluency

Previous studies of temporal phenomena in native and non-native speech have identified a number of quantitative variables that appear to be related to perceived fluency. In this context the term “temporal” does not refer exclusively to timing-related variables such as speaking rate, utterance duration, and pausing, but it also covers hesitation phenomena such as filled pauses, repetitions, and restarts (Grosjean, 1980).

Early studies of temporal phenomena were aimed at gaining more insight into psycholinguistic processes in one language (Goldman-Eisler, 1968). Subsequently, the analysis of temporal phenomena was applied in cross-linguistic investigations (Grosjean and Deschamps, 1975; Grosjean, 1980) and in studies of second language acquisition (Dechert and Raupach, 1980a, 1980b; Möhle, 1984). Recently, temporal variables have been employed in studies on perceived fluency and fluency development such as Nation (1989), Lennon (1990), Riggensbach (1991), Freed (1995), Towell *et al.* (1996).

On the basis of the literature on temporal variables in L2 acquisition and perceived fluency, the following measures were selected for investigation:

- (a) *ros* = rate of speech = # phonemes/total duration of speech including sentence-internal pauses
- (b) *ptr* = phonation/time ratio = 100% × total duration of speech without pauses/total duration of speech including sentence-internal pauses
- (c) *art* = articulation rate = # phonemes/total duration of speech without pauses

TABLE III. Intrarater and inter-rater reliability coefficients (Cronbach's  $\alpha$ ) for the three rater groups.

	Intrarater reliability			Inter-rater reliability	
	Rater 1	Rater 2	Rater 3	NNS & NS	NNS
Phoneticians	0.97	0.94	0.95	0.96	0.96
Speech therapists 1	0.94	0.97	0.96	0.93	0.88
Speech therapists 2	0.90	0.76	0.91	0.90	0.83

- (d)  $\#p$  = # of silent pauses = # of sentence-internal pauses of no less than 0.2 s
- (e)  $tdp$  = total duration of pauses = total duration of all sentence-internal pauses of no less than 0.2 s
- (f)  $mlp$  = mean length of pauses = mean length of all sentence-internal pauses of no less than 0.2 s
- (g)  $mlr$  = mean length of runs = average number of phonemes occurring between unfilled pauses of no less than 0.2 s
- (h)  $\#fp$  = # filled pauses = # of uh, er, mm, etc.
- (i)  $\#dy$  = # dysfluencies = # of repetitions, restarts, repairs

The first seven variables (*ros*, *ptr*, *art*, *tdp*,  $\#p$ , *mlp*, *mlr*) correspond to the Primary Variables in Grosjean's (1980) taxonomy, i.e., "variables that are always present in language output." The only differences are that we use phonemes as units instead of syllables and that we distinguish between number, total length, and mean length of silent pauses (see also Towell *et al.*, 1996). The latter two variables ( $\#fp$  and  $\#dy$ ) pertain to Grosjean's (1980) Secondary Variables, i.e., variables that are not necessarily present in speech. In addition, these variables seem to be infrequent in read speech (Grosjean, 1980), which would suggest that they are not good indicators of fluency in read speech. However, since it is not known how often they occur in read speech of non-natives, they are included in the present investigation.

In previous investigations, these variables were calculated manually (Möhle, 1984; Nation, 1989; Lennon, 1990; Riggenbach, 1991; Freed, 1995; Towell *et al.*, 1996), while in the present study the measures were calculated automatically by means of an automatic speech recognizer, as was explained in the previous section.

The various fluency scores for the individual sentences were subsequently averaged over the five sentences in each set and then over the two sets of each speaker. This way a set of 80 (60 NNS+20 NS) scores was obtained for each measure, which were then compared with the human-assigned fluency scores.

## II. RESULTS

In presenting the results of the present experiment, we will first pay attention to the expert fluency ratings. In particular, we will consider the issues of intrarater and inter-rater reliability. Subsequently, the relationship between the expert fluency ratings and the quantitative measures will be addressed. Finally, the differences between native and non-native speakers, both on the fluency ratings and on the quantitative measures, will be examined.

### A. Reliability of expert fluency ratings

The fluency ratings assigned by the three groups of experts were first analyzed to determine intrarater and inter-rater reliability. Intrarater reliability was calculated on the basis of  $12 \times 2$  scores for each rater, while the computation of inter-rater reliability was based on  $44 \times 3$  scores for each group of raters (44 sentence sets that were scored by all three raters in each group). The results of these analyses are shown in Table III.

As appears from Table III, intrarater reliability is very high for all raters, with the exception of rater 2 in the second group of speech therapists, who reaches only 0.76. Inter-rater reliability appears to be very high for all three groups. Since native 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, reliability was computed for two different conditions: (1) NS & NNS (both groups of speakers), (2) NNS (only foreign speakers). As is clear from Table III, even in the least favorable condition (NNS), the reliability coefficients are still rather high.

Besides considering inter-rater reliability, we also checked the degree of inter-rater 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 (see Table IV).

A low degree of 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 machine scores). This is so, because straightforward combination of the scores would amount to pooling measurements made with different yardsticks. When such a heterogeneous 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, 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, as is clear from Table IV.

The effect of normalizing the data is evident from Table V, which shows the correlation coefficients between the

TABLE IV. Means and standard deviations for the three raters in each group for the overlap material (the sentence sets used for determining inter-rater reliability) and for all the material scored by each rater.

		Rater 1		Rater 2		Rater 3	
		$\bar{x}$	sd	$\bar{x}$	sd	$\bar{x}$	sd
Phoneticians	overlap material	5.41	2.91	6.09	2.39	6.18	3.06
	all material	5.36	2.69	5.95	2.13	5.99	2.86
Speech therapists 1	overlap material	7.16	2.50	6.84	3.26	7.80	2.47
	all material	7.06	2.37	7.08	3.00	7.61	2.42
Speech therapists 2	overlap material	7.36	2.90	5.75	1.89	6.98	2.72
	all material	7.42	2.98	5.57	1.73	6.91	2.61

groups of raters before and after normalization. Since it is known that measurement errors affect the magnitude of the correlation coefficient, the correction for attenuation was applied (Ferguson, 1987), to allow direct comparisons between the various coefficients.

These correlations are so high that we can conclude that all nine raters involved in this experiment adopt similar definitions of fluency. Given the advantages of normalization, standard scores will be used also in the rest of the analyses in this study.

### B. Quantitative measures as indicators of perceived fluency

Before turning to the correlations among the fluency ratings and the temporal measures, we will first present the means and standard deviations of the nine temporal measures and the correlations among them.

The data in Table VI confirm that filled pauses and dysfluencies are indeed very infrequent in this type of speech. For this reason they will not be involved in the rest of the analyses presented in this paper. The mean value for articulation rate appears to be below the average of 15 phonemes per second indicated by Levelt (1989, p. 22) as average in normal speech. This is not surprising if we consider that these data refer to natives and non-natives and that articulation rate should be lower in non-natives (Towell *et al.*, 1996). Furthermore, since these data pertain to read speech, articulation rate should be lower than the average 15 phonemes per second also for native speakers. This point will be addressed in more detail in Sec. II C.

The correlations among the remaining seven quantitative variables are shown in Table VII. It is clear that all seven variables are relatively highly correlated with each other, but there are differences. For example, *ros*, *ptr*, *#p*, *tdp*, and *mlr* are highly correlated with each other ( $>0.86$ ). *art*, on the other hand, is highly correlated only with *ros*, while its correlations with the other variables are moderate (between 0.61 and 0.75). A clear exception is *mlp*, which shows moderate correlations with all other variables.

To establish which of the quantitative variables analyzed can be successfully used as a predictor of fluency in read speech, the correlations among the quantitative variables and the fluency ratings assigned by the experts were calculated. For the same reason as explained in Sec. III A, these correlations were calculated both for the whole group of speakers

(natives and non-natives) and for the non-natives only. The results of these analyses, corrected for attenuation, are shown in Table VIII.

From Table VIII it appears that all quantitative variables are strongly correlated with the fluency ratings, with the exception of *mlp*. For all three groups of raters, the highest correlation is found for *ros*. Moreover, it appears that the correlations for the non-natives are of the same order of magnitude as those for the whole group of speakers.

To determine whether a combination of variables allows us to make better predictions, we submitted these data to a multiple regression analysis in which the temporal variables are used as the predictors and the fluency ratings as the criterion. From Table V it appears that the fluency scores assigned by the three groups of raters are highly correlated with each other. For this reason we decided to use the mean scores in the regression analysis. The results of this analysis show that the variable that explains the greatest amount of variance is *ros*:  $R$  is 0.93. The second variable that is added in the stepwise procedure is *#p*. However, the increase in explained variance is marginal: Multiple  $R$  rises to 0.94.

### C. Differences between natives and non-natives

In this section we analyze both the fluency ratings and the seven quantitative measures to determine whether the two groups of natives and non-natives significantly differ on these variables. To this end, the two sets of data were submitted to a  $t$ -test for comparison of means. The results of these analyses are shown in Table IX. From this table it appears that the native speakers involved in this study were systematically found to be significantly 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. In addition, Table IX reveals that also for the native speakers in

TABLE V. Correlations among the groups of raters before and after normalization.

	Raw scores	Standard scores
Phoneticians-speech therapists 1	0.92	0.94
Phoneticians-speech therapists 2	0.82	0.90
Speech therapists 1-speech therapists 2	0.83	0.90

TABLE VI. Means and standard deviations for the nine quantitative variables.

	Rate of speech	Phonation/time ratio	Articulation rate	Number of pauses	Tot. duration of pauses	Mean length of pauses	Mean length of runs	Number of filled pauses	Number of dysfluencies
$\bar{x}$	10.44	85.29	12.12	5.76	2.43	0.33	24.71	0.11	0.49
sd	2.24	8.81	1.59	5.39	2.66	0.15	9.83	0.31	0.70

this experiment, articulation rate is indeed lower than the 15 phonemes per second indicated by Levelt (1989, p. 22) as average in normal speech.

Furthermore, Table IX shows that the native and the non-native speakers of Dutch in this study significantly differ from each other on all quantitative variables investigated. In other words, native speakers do appear to speak faster and to pause less than non-native speakers.

### III. DISCUSSION

In this paper we have presented the results of a study on perceived 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 quantitative measures that were automatically calculated for the same speech fragments. Reading material was purposely chosen in this study because it offers the possibility of reducing the impact of some linguistic factors known to affect fluency ratings (Riggenbach, 1991; Freed, 1995), while concentrating on the temporal variables as much as possible. A possible disadvantage of this choice is that it is not known whether the various degrees of fluency, or lack thereof, should be attributed to speech problems or to reading problems. However, if we consider that reading is often used in examinations in second or foreign language acquisition as a way of assessing fluency, then we may conclude that using read speech is less far-fetched than one might think at first.

The results of this study show that it is possible to obtain reliable ratings of fluency: reliability was high for all three groups of experts (Cronbach's  $\alpha$  varied between 0.90 and 0.96). 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 (Riggenbach, 1991; Freed, 1995). On the other hand, we had deliberately chosen read speech material so that the raters would be less distracted by other factors than those under study, as

explained above. In read speech, grammar and vocabulary can be kept constant. However, accent can still vary and can possibly affect the fluency ratings. In spite of this the raters achieved high reliability.

The major goal of this investigation was to determine whether automatically obtained quantitative measures of fluency can be used to predict expert fluency ratings. The results presented above show that automatic scoring of fluency in read speech is possible. As a matter of fact, six automatic measures showed correlations with the fluency scores which varied in magnitude between 0.81 and 0.93. *ros* appears to be the best predictor of perceived fluency (correlations vary between 0.90 and 0.93). According to the results of the regression analysis, the inclusion of other variables in the regression equation does not add much to the amount of explained variance, which is not surprising given that all variables are strongly correlated with each other (see Table VII) and that the correlations among *ros* and the fluency ratings are already so high. Moreover, it should be noted that the magnitude of the correlations among the fluency ratings and the temporal measures very much resembles those between the fluency ratings of the experts, which varied between .90 and 0.94 and which constitute a sort of upper bound for the man-machine correlations.

With respect to the contribution of the different variables to perceived fluency, Table VIII reveals that the fluency ratings are strongly affected by *ros*, *art*, *ptr*, *#p*, *tdp*, and *mlr*, while *mlp* has a smaller effect. This suggests that for perceived fluency the frequency of pauses is more relevant than their length. In other words, the difference between fluent and nonfluent speakers lies in the number of the pauses they make, rather than in their length, and the longer *tdp* of nonfluent speakers is caused by a greater number of pauses rather than by longer pauses. These findings are in line with those of previous investigations (see Chambers, 1997) and are corroborated by the analyses of the differences between natives and non-natives: Table IX shows that the

TABLE VII. Correlations among seven quantitative variables.

	Phonation/time ratio	Articulation rate	Number of pauses	Tot. duration of pauses	Mean length of pauses	Mean length of runs
Rate of speech	0.91	0.96	-0.87	-0.86	-0.71	0.88
Phonation/time ratio		0.75	-0.97	-0.96	-0.73	0.94
Articulation rate			-0.72	-0.71	-0.61	0.74
Number of pauses				0.97	0.63	-0.91
Tot. duration of pauses					0.67	-0.86
Mean length of pauses						-0.76

TABLE VIII. Correlations among the fluency ratings by the three rater groups and the quantitative measures, for the whole group ( $n=80$ ) and for the non-natives only ( $n=60$ ).

	Phoneticians		Speech therapists 1		Speech therapists 2	
	NNS & NS	NNS	NNS & NS	NNS	NNS & NS	NNS
Rate of speech	0.93	0.88	0.91	0.93	0.90	0.91
Phonation/time ratio	0.86	0.80	0.89	0.86	0.89	0.89
Articulation rate	0.88	0.82	0.85	0.86	0.81	0.79
Number of pauses	-0.84	-0.82	-0.89	-0.89	-0.89	-0.90
Tot. duration of pauses	-0.81	-0.79	-0.86	-0.86	-0.86	-0.87
Mean length of pauses	-0.66	-0.50	-0.62	-0.52	-0.65	-0.55
Mean length of runs	0.85	0.81	0.86	0.84	0.88	0.89

differences between natives and non-natives with respect to *mlp* are significant; however, these differences are relatively smaller than those concerning *#p* and *tdp*.

So, these results suggest that two important factors for perceived fluency in read speech are the rate at which speakers articulate the sounds and the number of pauses they make. *ros* appears to be such a good predictor of perceived fluency because it is a complex variable that incorporates the two aspects of articulation rate (number of segments) and pause time (*tdp*) (Chambers, 1997). *tdp* is of course dependent on the number of pauses, but the same *tdp* may be caused by a few long pauses or by many short pauses. In *ros*, this difference cannot be seen. In other words, although *ros* appears to be a very good predictor of reading fluency, it is possible that for certain purposes, for instance diagnostic ones, one may want to know how a specific score was obtained. In this case, adding the variable *#p* may be informative.

A possible limitation of these results is that they only indicate a strong relationship between objective measures of temporal speech characteristics on the one hand and expert fluency ratings on the other, but they do not provide information as to how varying articulation rate and/or pause time would affect the fluency ratings. In other words, we are not in a position to make strong claims about the causal relationships obtaining between the objective measures and the fluency ratings. One way of investigating this would be by compressing and expanding the speech under study, although this is not as simple as it might seem. Another possibility would be to use speech where a different relationship between articulation rate and pause time obtains, such as spon-

taneous speech. Since we are now working to extend the automatic approach to spontaneous speech, in the near future we will probably be able to address the issue of the causal relationship on the basis of spontaneous speech measurements. In any case, it is clear that this is a rather complex issue that deserves a series of studies on its own (see also, Butcher, 1981).

The results of this study indicate that automatically calculated temporal measures of speech could be used to develop objective tests of fluency, at least in read speech. In this sense this study is an answer to Lennon's call for more research along the lines of his own study, "but with larger sample groups" (Lennon, 1990), for "comparisons between learner and native-speaker performance" (Lennon, 1990), for "machine analysis of spoken text which...might be particularly useful when expert judges are not available to make an assessment" (Lennon, 1990) and "to develop standardized techniques for fluency assessment that would be independent of variation between individual raters" (Lennon, 1990). With respect to testing, however, it should be pointed out that in this study we were primarily exploring the possibilities of this approach and were not actually constructing a fluency test. This might explain why, for example, our focus was on reliability and less on agreement. In some cases agreement turned out not to be very high and we decided to use standard scores to combine the scores of the three raters in each group. The degree of agreement does play a crucial role in constructing a fluency test, because it contributes to establishing the cutoff point. However, since we are still in the development stage, agreement was less important in the present experiment, while reliability was our main concern.

TABLE IX. Results of *t*-tests for the fluency ratings of the three rater groups and for seven quantitative variables.

	<i>t</i> -test						
	$\bar{x}$ NS	sd NS	$\bar{x}$ NNS	sd NNS	<i>t</i> -value	df	<i>p</i>
Phoneticians	0.88	0.39	-0.32	0.70	9.55	59.98	0.000
Speech therapists 1	0.91	0.13	-0.27	0.79	11.07	67.55	0.000
Speech therapists 2	0.86	0.33	-0.30	0.83	8.90	75.77	0.000
Rate of speech	12.74	1.35	9.68	1.94	6.54	78	0.000
Phonation/time ratio	93.17	2.79	82.66	8.57	8.27	78	0.000
Articulation rate	13.65	1.19	11.61	1.37	5.97	78	0.000
Number of pauses	1.42	1.23	7.20	5.47	-7.62	73	0.000
Tot. duration of pauses	0.45	0.42	3.10	2.76	-7.18	66.68	0.000
Mean length of pauses	0.20	0.13	0.38	0.13	-5.24	78	0.000
Mean length of runs	34.26	5.85	21.52	8.77	7.36	49.2	0.000

The potential of this approach for automatic fluency assessment is all the more important if we consider that these results pertain to telephone speech. 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.

The data collected in this study were also analyzed to determine whether the two groups of native and non-native speakers significantly differ on perceived fluency and on seven quantitative measures of fluency. The results reveal significant differences between the two groups on all variables. As mentioned above, these results indicate that natives and non-natives are more different from each other with respect to pause frequency than to pause length. Furthermore, 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 (Riggenbach, 1991), these results show that, on average, they are more fluent, produce fewer pauses, and articulate faster than non-native speakers.

#### IV. CONCLUSIONS

On the basis of the results of the present investigation we can draw the following conclusions. First, expert listeners are able to evaluate fluency with a high degree of reliability. Second, expert fluency ratings of read speech are mainly influenced by two factors: speed of articulation and frequency of pauses. Third, expert fluency ratings can be accurately predicted on the basis of automatically calculated measures such as rate of speech, articulation rate, phonation–time ratio, number and total duration of pauses, and mean length of runs. Of all these measures, rate of speech appears to be the best one. Fourth, native speakers are more fluent than non-natives and the temporal measures are significantly different for the two groups.

To conclude, these findings indicate that temporal measures of fluency may be employed to develop objective testing instruments of fluency in read speech. In turn, the fact that these measures can be automatically calculated by means of automatic speech recognition techniques suggests that this approach 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, then it seems that this approach is likely to have important consequences for the future of fluency assessment.

#### ACKNOWLEDGMENTS

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 Zeitlinger, and KPN. The research of Dr. H. Strik has been made possible by a fellowship of the Royal Netherlands Academy of Arts and Sciences. The authors would like to thank Tim Bunnell and an anonymous reviewer for their valuable comments and suggestions.

#### APPENDIX

##### Group 1 sentences

- (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.

##### Group 2 sentences

- (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.

Brumfit, C. (1984). *Communicative Methodology in Language Teaching: The Roles of Fluency and Accuracy* (Cambridge University Press, Cambridge).

Butcher, A. (1981). “Phonetic correlates of perceived tempo in reading and spontaneous speech,” *Work in Progress*, Phon. Lab. Univ. Reading, pp. 105–117.

Chambers, F. (1997). “What do we mean by fluency?,” *System* 4, 535–544.

Dechert, H. W., and Raupach, M. (1980a). *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler* (Mouton, The Hague).

Dechert, H. W., and Raupach, M. (1980b). *Towards a Cross-Linguistic Assessment of Speech Production* (Lang, Frankfurt).

den Os, E. A., Boogaart, T. I., Boves, L., and Klabbers, E. (1995). “The Dutch Polyphone Corpus,” *Proceedings Eurospeech95*, pp. 825–828.

Ferguson, G. A. (1987). *Statistical Analysis in Psychology and Education* (McGraw-Hill, Singapore).

Fillmore, C. J. (1979). “On fluency,” in *Individual Differences in Language Ability and Language Behavior*, edited by C. Fillmore, D. Kempler, and W. S.-Y. Wang (Academic, New York), pp. 85–101.

Flege, J. E., and Fletcher, K. L. (1992). “Talker and listener effects on degree of perceived foreign accent,” *J. Acoust. Soc. Am.* 91(1), 370–389.

Freed, B. F. (1995). “What makes us think that students who study abroad become fluent?,” in *Second Language Acquisition in a Study-Abroad Context*, edited by B. F. Freed (Benjamins, Amsterdam), pp. 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*, edited by H. W. Dechert and M. Raupach (Lang, Frankfurt), pp. 39–53.

- Grosjean, F., and Deschamps, A. (1975). "Analyse contrastive des variables temporelles de l'Anglais et du Français: Vitesse de parole et variables composantes, phénomènes d'hésitation," *Phonetica* **31**, 144–184.
- Leeson, R. (1975). *Fluency and Language Teaching* (Longman, London).
- Lennon, P. (1990). "Investigating fluency in EFL: A quantitative approach," *Language Learning* **3**, 387–417.
- Levelt, W. J. M. (1989). *Speaking. From Intention to Articulation* (MIT Press, Cambridge, MA).
- Möhle, D. (1984). "A comparison of the second language speech production of different native speakers," in *Second Language Productions*, edited by H. W. Dechert, D. Möhle, and M. Raupach (Narr, Tübingen), pp. 26–49.
- Nation, P. (1989). "Improving speaking fluency," *System* **3**, 377–384.
- Peters, T. J., and Guitar, B. (1991). *Stuttering. An Integrated Approach to its Nature and Treatment* (William and Wilkins, Baltimore).
- Raupach, M. (1980). "Temporal variables in first and second language speech production," in *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*, edited by H. W. Dechert and M. Raupach (Mouton, The Hague), pp. 263–270.
- Raupach, M. (1983). "Analysis and evaluation of communicative strategies," in *Strategies in Interlanguage Communication*, edited by C. Faerch and G. Kasper (Longman, London), pp. 263–270.
- Riggenbach, H. (1991). "Toward an understanding of fluency: A microanalysis of non-native speaker conversations," *Discourse Process* **14**, 423–441.
- Schmidt, R. (1992). "Psychological mechanisms underlying second language fluency," *Stud. Second Language Acquisition* **14**, 357–385.
- Segalowitz, N. (1991). "Does advanced skill in a second language reduce automaticity in the first language?," *Language Learning* **41**, 59–83.
- SPEX <http://lands.let.kun.nl/spex>.
- Strik, H., Russel, A., Van den Heuvel, H., Cucchiari, C., and Boves, L. (1997). "A spoken dialog system for the Dutch Public Transport Information Service," *International J. Speech Technol.* **2**, 121–131.
- Towell, R. (1987). "Approaches to the analysis of the oral language development of the advanced learner," in *The Advanced Language Learner*, edited by J. A. Coleman and R. Towell (CILT, London), pp. 157–181.
- Towell, R., Hawkins, R., and Bazergui, N. (1996). "The development of fluency in advanced learners of French," *Appl. Linguistics* **1**, 84–119.