Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech

Catia Cucchiarini,[a] Helmer Strik, and Lou Boves
A^2RT, Department of Language and Speech, University of Nijmegen, P.O. Box 9103, 6500 HD Nijmegen, The Netherlands

(Received 2 February 2001; revised 23 October 2001; accepted 15 February 2002)

This paper describes two experiments aimed at exploring the relationship between objective properties of speech and perceived fluency in read and spontaneous speech. The aim is to determine whether such quantitative measures can be used to develop objective fluency tests. Fragments of read speech (Experiment 1) of 60 non-native speakers of Dutch and of spontaneous speech (Experiment 2) of another group of 57 non-native speakers of Dutch were scored for fluency by human raters and were analyzed by means of a continuous speech recognizer to calculate a number of objective measures of speech quality known to be related to perceived fluency. The results show that the objective measures investigated in this study can be employed to predict fluency ratings, but the predictive power of such measures is stronger for read speech than for spontaneous speech. Moreover, the adequacy of the variables to be employed appears to be dependent on the specific type of speech material investigated and the specific task performed by the speaker. © *2002 Acoustical Society of America.* [DOI: 10.1121/1.1471894]

PACS numbers: 43.71.Es, 43.71.Gv, 43.71.Hw, 43.72.Ne [DOS]

I. INTRODUCTION

Oral fluency is viewed as an important characteristic of second language speech, which explains why it is often the object of evaluation in testing second language skills (Riggenbach, 1991; Freed, 1995). In spite of the wide use that is made of this notion with respect to second language speech, there is no generally agreed definition of fluency and this term has been used to indicate different constructs. In everyday language use, fluency is often considered as a synonym of ''overall language performance.'' In this interpretation native-like performance is viewed as the final goal. Brumfit (1984, p. 57), on the other hand, defines fluency as ''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). A different, more restricted definition of fluency is the one that refers to the temporal aspect of speech (Nation, 1989; Lennon, 1990; Riggenbach, 1991; Schmidt, 1992; Freed, 1995; Towel *et al.*, 1996) and puts emphasis on ''native-like rapidity'' (Lennon, 1990, p. 390). According to this interpretation, the goal in second language learning would be to produce ''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, p. 390). However, various quantitative studies have revealed that even native speech, far from being always smooth and continuous, exhibits many hesitations and repairs (Raupach, 1983; Lennon, 1990; Riggenbach, 1991).

With a view to gaining more insight into the factors that contribute to perceived fluency, attempts have been made to try to define fluency in terms of objective properties of speech (Lennon, 1990; Riggenbach, 1991; Freed, 1995). These studies have adopted a dual approach in which listeners' evaluations of speech, in this case perceived fluency scores, are related to objective measures calculated for the same speech. This type of approach, which proved particularly useful to gain insight into the dimensions underlying the listeners' evaluations, has a long tradition in phonetic research and has been used previously in other domains such as the evaluation of voice (Laver, 1980), running speech (Boves, 1984), vocal expressions of emotions (Van Bezooijen, 1984), and non-native pronunciation (Neumeyer *et al.*, 1996). Another important aspect of fluency studies based on this method is that they may contribute to developing more objective and, possibly, less labor intensive tests of second language fluency (see, for instance, Townshend *et al.*, 1998). With the growing numbers of immigrants who have to learn other languages, the practical advantages of objective fluency measures are obvious.

In Cucchiarini *et al.* (2000) a study was described in which the dual approach was adopted to gain insight into the temporal definition of fluency. That study differed from previous ones in two important respects. First, the objective measures, which were calculated manually in previous studies (Lennon, 1990; Riggenbach, 1991; Freed, 1995), were calculated automatically with the help of a continuous speech recognizer (CSR), which has the advantage that huge amounts of data can be analyzed in relatively short time and in a very consistent manner. Second, instead of using spontaneous speech, read speech was used, so that the raters would not be distracted by differences in vocabulary and syntax-related parameters, which are known to affect fluency ratings (Riggenbach, 1991; Freed, 1995). However, our idea was to apply this approach to spontaneous speech too, if it turned out to be feasible for read speech.

The experiment reported in Cucchiarini *et al.* (2000)

[a] Electronic mail: c.cucchiarini@let.kun.nl

produced interesting results in two respects. First, the results showed that the expert ratings of fluency in read speech were reliable (Cronbachs' α varied between 0.90 and 0.96), which contrasted with the much lower reliability coefficients reported in previous studies (Riggenbach, 1991; Freed, 1995), but seemed plausible given that those studies concerned spontaneous speech. Second, very high correlations were found between the expert fluency ratings and the objective measures of fluency: five objective measures showed correlations with the fluency scores with magnitudes between 0.77 and 0.91. 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 an excellent predictor of perceived fluency because it incorporates these two aspects.

The results of this previous study on read speech also raised some questions, the most important being: would these results hold for spontaneous speech too? In particular, it seemed interesting to determine whether the quantitative measures that were found to affect perceived fluency in read speech would be equally important for perceived fluency in spontaneous speech and/or, the other way round, whether there are measures that are suitable for spontaneous speech, but not for read speech. In an attempt to find answers to these questions, an experiment with spontaneous speech was carried out in which the dual approach was used to investigate the same temporal notion of fluency.

As explained previously, the importance of this type of research is not only related to the possibility of getting more insight into the relationship between perceived fluency and temporal characteristics of speech, but also to the potential that this kind of research might have for the development of objective testing instruments for fluency assessment, especially in the context of second language teaching and testing. Against this background it seemed more advantageous to use an existing test of second language proficiency rather than collecting speech material especially for this experiment. In this way the material under study would be less of the ''laboratory'' type and would be more similar to what is generally found in the ''field.'' On the one hand, this might have the disadvantage that the experimenter cannot control all aspects of the experiment. On the other hand, it has the considerable advantage that in this way external validity is guaranteed. Since it is clear that the importance of external validity cannot be overestimated in these kinds of studies and that the advantages of using a real test evaluated by real raters outweigh the disadvantages of using a less elegant experimental design, it was decided to use an already existing test of second language proficiency that would be suited for our purpose.

The test that was eventually selected for this experiment is the *Profieltoets*. This test was developed by the Dutch National Institute for Educational Measurement (Cito) and is normally administered to immigrants who, within the framework of the Newcomer Integration Act, are obliged to follow a Dutch language course upon arrival in The Netherlands. In general, newcomers take the *Profieltoets* after they have followed about 500–600 h of lessons in the Dutch language. In this test the four skills speaking, listening, reading, and writing are tested. For this experiment a subtest corresponding to the speaking test was used. This test is administered in a language lab to a group of several candidates simultaneously. The candidates have to answer questions which elicit unprepared answers. The speech can therefore be classified as extemporaneous, spontaneous speech. In other words, this experiment does not concern speech that was especially elicited for the purpose of the experiment, but speech that subjects produced while they were taking a real examination.

As in our previous study, the dual approach was adopted in which the speech material was evaluated by a group of raters and by an automatic speech recognizer which was used to calculate a number of objective measures that are known to be related to perceived fluency. The aim of the present paper is to explore the relationship between these objective properties of speech and perceived fluency in read and spontaneous speech, with a view to determining whether such quantitative measures can be used to develop objective fluency tests. To pursue this aim, the read speech data from our previous experiment were compared with the spontaneous speech data from the present one. In the read speech experiment the 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. Since in the spontaneous speech experiment only non-natives were involved (recall that these data stemmed from a real second language proficiency test), from the read speech experiment only the data pertaining to the 60 non-native speakers were used. These two experiments will be referred to as Experiment 1 (read speech) and Experiment 2 (spontaneous speech). Although Experiment 1 has already been presented in detail in Cucchiarini *et al.* (2000), the data concerning the non-native speakers were not presented as explicitly as they are in this paper. In any case, here we will limit ourselves to providing only the Experiment 1 data and details that are necessary to make comparisons between read speech (Experiment 1) and spontaneous speech (Experiment 2) of non-native speakers of Dutch.

II. METHOD

A. Speakers

1. Experiment 1

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

2. Experiment 2

The speakers involved in this experiment constitute a subgroup of the candidates who took part in the test *Profieltoets* in June 1998. In this investigation the answers of 60 subjects of two different proficiency levels were analyzed: a lower proficiency group at the beginner level (BL) and a higher proficiency group at the intermediate level (IL). These two proficiency levels roughly correspond to the first two levels (PL1 and PL2) in Experiment 1. Cito workers selected for us a subgroup of 30 speakers at the beginner level and a subgroup of 30 speakers at the intermediate level. In both groups the speakers varied with respect to gender and mother tongue.

B. Speech material

1. Experiment 1

Each speaker read two different sets of five phonetically rich sentences designated set 1 and set 2, respectively. The only difference between sets 1 and 2 is that they contain different sentences. However each group of five sentences contains all phonemes of Dutch at least once, while more common phonemes appear more than once. Since the average duration of each set was 30 s, almost 1 min of speech per speaker was available. The sentences were printed on paper together with the instructions and were read by the speakers over the telephone. The subjects had not explicitly been encouraged to rehearse before reading, but since they had received the material beforehand, they had this possibility. They also had the possibility of restarting the recording session 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 consisted of 8 kHz, 8 bit, A-law coded samples. Almost all subjects called from their homes, while two called from the first author's office. No provisions were made to control background noise; consequently, the acoustic quality of the recordings varied considerably.

2. Experiment 2

The speech material used in this experiment consisted of the answers given by the candidates mentioned in Sec. II A 2 to part of the items which constitute the *Profieltoets*. This test is administered to candidates at the beginner and intermediate proficiency levels and is therefore available in two different versions corresponding to the two groups, BL and IL. For this experiment eight items were selected for each version of the test. The items differed for the two proficiency groups, because in this case we have less influence on the selection of the material. This is a consequence of choosing an existing test. An important requirement in selecting the items was that they had to elicit relatively long answers, which is a necessary condition for calculating fluency measures.

For the IL group the so-called long tasks were chosen, in which the candidates have 30 s to answer each question. In these items the candidate has to answer questions by choosing from among various possibilities and has to explain why he/she made that choice. In other words, the candidate, when answering, has to reflect to find good motivations for his/her choice.

The BL version of the test contains only the short tasks, in which the subjects have 15 s at their disposal to answer each question. In general, in these items a given situation is presented and the candidate has to indicate what he/she would say in that context. From these tasks we chose those for which, given the nature of the questions, reasonably long answers of at least a few words can be expected. Effectively, the BL subjects talked for about 70 s on average, while for the IL subjects the average was 180 s.

The fact that the BL items elicit rather straightforward responses, whereas the IL items seem to require more cognitive effort could have an impact on the fluency scores, as has been reported by Grosjean (1980) and Bortfeld *et al.* (1999). In particular, Grosjean (1980, pp. 42–43) has shown that more cognitively demanding tasks will lead to a lower speech rate, which, in turn, is accounted for by a lower articulation rate, but especially by a lower phonation/time ratio, shorter runs, and longer pauses.

Finally, it should be observed that the speech samples thus obtained are different for the various speakers, an obvious implication of using extemporaneous speech, whereas in Experiment 1 all speakers produced the same sentences, which is easy to achieve with read speech.

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

C. Raters

1. Experiment 1

Since previous studies had revealed that expert fluency ratings displayed low reliability (Lennon, 1990; Riggenbach, 1991; Freed, 1995), in this experiment it was decided to ask multiple groups of experts to evaluate the speech material: a group of three phoneticians (PH) and two groups of three speech therapists (ST1 and ST2). The phoneticians were chosen for obvious reasons, whereas the speech therapists were chosen because their expertise is usually invoked when learners of Dutch exhibit pronunciation problems, including all fluency-related temporal phenomena. As reported in Cucchiarini *et al.* (2000), reliability appeared to be very high for all three groups of raters.

2. Experiment 2

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. To be able to work as raters for Cito these teachers have to take a three-day course which they have to conclude with an examination. Furthermore, their performance as raters in different kinds of tests administered by Cito is regularly checked and Cito workers keep track of each rater's performance by calculating overall indices of reliability that can be made available if required.

The scoring sessions were organized by Cito according to the procedure that is usually followed for the *Profieltoets*. One group of five teachers, designated as raters for the beginning level (RBL), evaluated the BL speakers and another group of five teachers designated as raters for the intermediate level (RIL), evaluated the IL speakers. There was no overlap of speakers between the two rater groups, which means that there was no subgroup of speakers that was evaluated by both rater groups.

D. Fluency ratings

1. Experiment 1

The speech material was transferred from disk to DAT tape adopting different random orders for the different raters. All raters listened to the speech material and evaluated perceived fluency 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 set 1 and the set 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 speakers were proportionally assigned to the three raters in each group. However, part of the material (overlap material) was scored by all three raters in one group so as to allow reliability checks. For further details on this point, see Cucchiarini *et al.* (2000).

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 60 human-assigned fluency scores were obtained, which were subsequently compared with the various quantitative measures.

2. Experiment 2

All raters listened to the speech material on audio cassettes and assigned scores individually. The raters first scored each speaker on the *Profieltoets* as they normally do. Subsequently, they were asked to score the eight selected items on fluency. The raters could listen to the speech fragments as often as they wanted. They were asked to score fluency on a scale ranging from 1 to 10. Each rater assigned one fluency score per set of eight items so that for each speaker in this experiment five fluency scores assigned by five raters were obtained. As in the experiment in Cucchiarini *et al.* (2000), no specific instructions were given for fluency assessment, but, as mentioned previously, these raters had all received a three-day training before starting to work as raters for Cito.

3. Experiment 1 versus Experiment 2

Two important differences between the two experiments should be mentioned. First, in Experiment 2 the two groups of speakers, BL and IL, were assigned to two different groups of raters, RBL and RIL, whereas in Experiment 1 each group of three raters evaluated all 60 speakers. This point should be borne in mind because it has consequences for the analyses that can be carried out and for the results of these experiments.

Second, the phoneticians and speech therapists involved in Experiment 1 simply judged the speech of a number of speakers without having information on the proficiency level of each speaker, except the cues that they could derive from the speech itself. The language teachers in Experiment 2, on the other hand, were judging candidates in an examination and therefore knew whether a speaker was at the beginner or intermediate level. As a consequence, they may have judged fluency in relation to each speaker's assumed proficiency level, so that the same score, say eight, would not have the same meaning in the two groups, but would represent a higher fluency level in the IL group than in the BL group.

E. Objective assessment of fluency

This part of the analysis procedure was the same for Experiments 1 and 2. All speech material was orthographically transcribed by SPEX (http://www.spex.nl/), an expertise center that specializes in database construction and validation. The material was also checked on quality both by SPEX and the first author (C.C.) before being used for the experiment. The recordings of three speakers in Experiment 2 could not be used because their quality was so poor that they were not even scored by the raters. This way a total of 28 BL speakers and 29 IL speakers was obtained.

In transcribing the material, special symbols were used for four categories of nonspeech acoustic events (as is usually done at SPEX):

(1) filled pauses: uh, er, mm, etc.
(2) speaker noise: lip smack, throat clear, tongue click, etc.
(3) intermittent noise: noise that occurs incidentally during the call such as door slam and paper rustle.
(4) 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 and were indicated by a special disfluency symbol so that they could be counted automatically.

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 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 filter band 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 context-independent 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, since a distinction was made between prevocalic (/l/ and /r/) and postvocalic position (/L/ and /R/). For each of the other 33 phonemes of Dutch, 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; each set of five sentences contains 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 CSR was subsequently used to analyze the utterances produced by the speakers. For each utterance a Viterbi alignment between the speech signal and the canonical phonemic transcription, which was generated automatically from the orthographic transcription, was obtained. For the purpose of the research in this paper only the boundaries between speech and nonspeech signals (silences, but also filled pauses) are relevant. The accuracy of forced alignment was checked manually for a representative sample of the material. In general, the segmentation appeared to be of sufficient quality for this purpose and was then used to calculate the quantitative measures which are described in detail in the following.

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 (Goldman-Eisler, 1968; Grosjean and Deschamps, 1975; Grosjean, 1980; Nation, 1989; Lennon, 1990; Riggenbach, 1991; Freed, 1995; Towell *et al.*, 1996). The clearest taxonomy is provided by Grosjean (1980, p. 40), who distinguishes between primary and secondary variables. Primary variables are ‘‘variables that are always present in language output’’ (Grosjean, 1980, p. 40). Secondary variables are related to hesitation phenomena such as filled pauses, repetitions, repairs, and restarts. These variables are not necessarily present in speech and seem to be infrequent in read speech (Grosjean, 1980, p. 42).

Before introducing the variables used, we first give some definitions:

(1) silence: every frame of silence detected by the CSR,
(2) silent pause: a stretch of silence with a duration of no less than 0.2 s,
(3) nph: number of phonemes,
(4) dur1: duration of speech without utterance internal silences,
(5) dur2: duration of speech including utterance internal silences.

TABLE I. Definition of quantitative fluency measures. dur1 = duration of speech without utterance internal silences, dur2 = duration of speech including utterance internal silences.

Name	Definition
Seven primary variables	
Articulation rate	Number of phonemes/dur1
Rate of speech	Number of phonemes/dur2
Phonation/time ratio	100% × dur1/dur2
Mean length of runs	Mean number of phonemes between silent pauses
Mean length of silent pauses	Mean length of all silent pauses
Duration of silent pauses per minute	Total duration of all silent pauses/(dur2/60)
Number of silent pauses per minute	Number of silent pauses/(dur2/60)
Two secondary variables	
Number of filled pauses per minute	Number of filled pauses/(dur2/60)
Number of disfluencies per minute	Number of disfluencies/(dur2/60)

dur1 and dur2 were measured from the beginning of the first word to the end of the last word for every utterance. Consequently, silences present at the beginning and end of every utterance were discarded.

These definitions were employed to calculate seven primary and two secondary variables, i.e., primary variables are further divided into complex variables and simple variables. Complex variables are speaking rate and phonation time ratio, while simple variables are articulation rate, length of silent pauses, and length of runs. As is clear from Table I, the Simple Variables are subcomponents of the Complex Variables. In our previous study (Cucchiarini *et al.*, 2000) as well as in the present one, measures similar to those proposed by Grosjean (1980) were adopted, albeit with slightly different definitions. These variables differ from those defined by Grosjean (1980) in three respects. First, phonemes were used as units instead of syllables. Second, a distinction was made between mean length, total length, and number of silent pauses (see also Towell *et al.*, 1996). Third, the variables *duration of silent pauses per minute, number of silent pauses per minute, number of filled pauses per minute, and number of disfluencies per minute* are all calculated relative to utterance length (i.e., dur2/60). This was not done in our previous paper (Cucchiarini *et al.*, 2000), because in that case all speakers read the same sentences, and thus the utterances of different subjects had almost the same length. In the present study, on the other hand, speech fragments of different length have to be compared and such absolute measures are not suitable for this purpose. As time unit the minute was chosen because it seemed most appropriate to express the frequency of phenomena such as filled and silent pauses and disfluencies. As indicated by Grosjean, disfluencies and filled pauses appeared to be infrequent in read speech (Cucchiarini *et al.*, 2000). However, it was decided to include them in the present investigation because they could be more frequent in spontaneous speech, especially that of non-natives.

TABLE II. Interrater reliability coefficients (Cronbach's α) for the five rater groups.

Rater group	Interrater reliability
Phoneticians	0.96
Speech therapists 1	0.88
Speech therapists 2	0.83
Raters beginner level	0.86
Raters intermediate level	0.82

For each speaker in Experiment 1 the above-mentioned objective fluency measures were calculated over the five sentences in each set, thus obtaining two values per objective measure for each speaker. These two values were then averaged so as to obtain one value per objective measure per speaker. With 60 speakers a total of 60 values was obtained for each objective measure. In Experiment 2 the objective fluency measures were calculated over the eight items of each speaker, thus obtaining a set of 57 values for each objective measure. In this manner two sets of 60 (Experiment 1) and 57 (Experiment 2) scores were obtained for each objective measure. Correlations between these values and the human-assigned fluency scores were then calculated.

III. RESULTS

In presenting the results of the two experiments, attention is first paid to the fluency ratings assigned by the various groups of raters. Subsequently, the results concerning the quantitative measures of fluency are examined. Finally, the relationship between the human-assigned fluency ratings and the quantitative measures is considered.

A. Fluency ratings

The fluency scores assigned by the various rater groups involved in the two experiments, PH, ST1, and ST2 for Experiment 1 and RBL and RIL for Experiment 2, were analyzed to determine interrater reliability. The results of these analyses are shown in Table II. As is clear from Table II, interrater reliability is reasonably high, which may be surprising in view of the low reliability coefficients reported by Lennon (1990), Riggenbach (1991), and Freed (1995).

Besides considering interrater reliability, we also checked the degree of interrater agreement. Close inspection of the data revealed that in both experiments the means and standard deviations varied between the various raters. In other words, in both experiments the raters differed from each other in degree of strictness. As a consequence, the degree of agreement was not very high. A low degree of agreement within a group of raters has the effect of lowering 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). 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, before calculating the correlation coefficients between the human-assigned fluency ratings and the objective measures it was necessary to normalize for the differences in the values by using standard scores instead of raw scores. In Experiment 1, the scores were normalized by using 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 mean and standard deviations for the total material, as was illustrated in Cucchiarini *et al.* (2000). In Experiment 2 normalizing the scores was more straightforward, because all five raters in one group rated all speakers. For each rater his/her mean was then subtracted from each of his/her scores and the resulting scores were then divided by the standard deviation for that rater.

Table III shows the mean and standard deviations (raw scores) of the fluency ratings for the speakers in the two experiments. In Table III it can be clearly seen that the read speech fluency scores vary for the three proficiency levels PL1 (beginner), PL 2 (intermediate), and PL3 (advanced) and that they gradually increase from PL1 to PL3, which means that the more proficient speakers are also perceived as being more fluent than the less proficient speakers. In the spontaneous speech data this relationship between proficiency and fluency does not seem to obtain, as the scores for the IL speakers are lower than those for the BL speakers. Although one might argue that the scores for the two speaker groups are not really comparable because they were assigned by two different groups of raters, it seems that these results are probably related to the context within which the evaluation was carried out.

As explained previously, the raters in Experiment 1 had no information about the proficiency level of each speaker, except the cues contained in the speech, whereas the raters in Experiment 2 knew to which proficiency group the speaker belonged. As a consequence, they judged fluency in relation to each speaker's proficiency level, thus assigning higher scores to less proficient speakers if the desired fluency level was lower, i.e., in the BL group. Another possibility is that the failure to find the expected difference between the two proficiency groups in Experiment 2 is due to the difference between the tasks performed by the two groups. As explained previously, the IL group carried out cognitively more demanding tasks which might have induced lower fluency

TABLE III. Means and standard deviations for the raw fluency scores for read and spontaneous speech of speakers of different proficiency levels.

	Read speech (RS)								Spontaneous speech (SS)					
	PL1		PL2		PL3		all-RS		BL		IL		all-SS	
	$\overline{x}$	s.d.	$\overline{x}$	s.d.	$\overline{x}$	s.d.	$\overline{x}$	s.d.	$\overline{x}$	s.d.	$\overline{x}$	s.d.	$\overline{x}$	s.d.
Raw fluency scores	4.65	2.01	5.00	1.81	7.36	0.95	5.85	1.96	5.64	0.88	4.80	1.06	5.21	1.06

TABLE IV. Means and standard deviations (in parentheses) for the quantitative measures for read and spontaneous speech of speakers of different proficiency levels, PL1 (beginner), PL2 (intermediate), PL3 (advanced), BL (beginner level), and IL (intermediate level). For the various subgroups *N* is indicated in square brackets.

	Read speech (RS)				Spontaneous speech (SS)			9
Column 1	2 PL1 [10] $\overline{x}$ (s.d.)	3 PL2 [27] $\overline{x}$ (s.d.)	4 PL3 [23] $\overline{x}$ (s.d.)	5 all-RS [60] $\overline{x}$ (s.d.)	6 BL [28] $\overline{x}$ (s.d.)	7 IL [29] $\overline{x}$ (s.d.)	8 all-SS [57] $\overline{x}$ (s.d.)	SS/RS ratio col. 8 col. 5
nph	430.0 (7.0)	434.6 (16.7)	428.8 (7.5)	431.6 (12.6)	417.4 (115.1)	950.2 (246.9)	688.4 (330.3)	
dur1	44.5 (5.1)	43.9 (5.8)	38.3 (2.8)	41.9 (5.5)	34.6 (10.6)	80.3 (20.1)	57.9 (28.1)	
dur2	57.9 (10.3)	56.4 (12.7)	43.7 (5.2)	51.8 (11.8)	70.0 (16.0)	179.9 (29.8)	125.9 (60.3)	
Articulation rate	10.87 (1.41)	11.15 (1.38)	12.47 (0.82)	11.61 (1.37)	12.25 (1.25)	11.85 (0.81)	12.04 (1.06)	1.0
Rate of speech	8.54 (1.88)	8.95 (1.87)	11.03 (1.16)	9.68 (1.94)	5.99 (0.96)	5.31 (1.17)	5.65 (1.12)	0.6
Phonat./time ratio	77.97 (7.69)	79.62 (8.68)	88.28 (5.42)	82.66 (8.57)	49.33 (8.71)	44.92 (9.51)	47.09 (9.32)	0.6
$\overline{x}$ length of runs	16.51 (7.67)	18.10 (7.44)	27.73 (7.13)	21.5 (8.77)	9.50 (2.22)	9.33 (2.27)	9.41 (2.23)	0.4
$\overline{x}$ length sil. paus.	0.40 (0.08)	0.40 (0.12)	0.34 (0.16)	0.38 (0.13)	0.92 (0.20)	1.02 (0.28)	0.97 (0.25)	2.6
Dur. sil. paus. p/m	9.29 (4.48)	8.67 (5.15)	3.97 (2.96)	6.97 (4.87)	27.90 (5.52)	31.02 (6.04)	29.49 (5.95)	4.2
No. sil. paus. p/m	22.33 (8.45)	20.11 (9.45)	10.18 (6.45)	16.67 (9.65)	31.00 (5.56)	31.41 (4.77)	31.21 (5.13)	1.9
No. fil. paus. p/m	0.31 (0.50)	0.35 (0.94)	0.32 (0.77)	0.33 (0.81)	10.83 (8.24)	10.55 (7.84)	10.69 (7.97)	32.4
No. disf. p/m	1.82 (2.20)	1.78 (1.75)	1.04 (1.53)	1.50 (1.76)	2.39 (1.82)	2.19 (2.27)	2.29 (2.04)	1.5

scores. The analyses of the quantitative fluency measures will shed light on this point.

B. Quantitative measures of fluency

In this section the quantitative variables are analyzed in various respects. First, the mean and standard deviation are calculated for all variables for all groups. These results are given in Table IV.

The rows nph, dur1, and dur2 give an indication of the amount of material that was analyzed for the current study on read speech (RS) and spontaneous speech (SS).

Table IV also shows how the values for the different variables vary as a function of speech modality (read versus spontaneous) and proficiency level. In order to see how the measures vary as a function of speech modality the means for read speech (column 5) can be compared with those pertaining to spontaneous speech (column 8). In order to facilitate this comparison, the SS/RS ratio for the seven primary and the two secondary variables was calculated by dividing the averages in column 8 by the averages in column 5. The results are presented in column 9.

These comparisons indicate that for most of the variables the values drastically change as we go from read speech to spontaneous speech. For the primary variables, *rate of speech, phonation/time ratio, and mean length of runs* are roughly halved (SS/RS ratio: 0.6, 0.6, and 0.4, respectively), *number of silent pauses per minute* is almost doubled (1.9), while *mean length of silent pauses* and *duration of silent pauses per minute* are more than doubled (2.6 and 4.2, respectively). On the other hand, *articulation rate* hardly changes (1.0).

As to the secondary variables, disfluencies are somewhat more frequent in spontaneous speech as compared to read speech (1.5), whereas the frequency of filled pauses is more than 30 times higher in spontaneous speech (32.4). Furthermore, it is clear that for the secondary variables the value of the standard deviation is relatively high with respect to the mean. In some cases the standard deviation is even much higher than the mean. This means that, instead of being monomodal, the frequency distributions of disfluencies and filled pauses are mainly characterized by extremely low and extremely high values.

In order to see how the quantitative measures vary as a function of proficiency level, we can compare columns 2, 3, and 4 within read speech and columns 6 and 7 within spontaneous speech. The first thing to be observed is that the values change as a function of proficiency level. In the read speech material gradual changes can be observed for the primary variables from PL1 to PL3. The change is either an increase or a decrease, depending on the variable in question, but all changes indicate that the less proficient speakers also obtain lower fluency scores in terms of these quantitative measures. In the spontaneous speech material the opposite seems to hold: the values of the primary variables for the less proficient speakers indicate higher fluency than those of the more proficient speakers. This is all the more remarkable because it holds for all seven measures. On the one hand, these results are in line with those presented in the previous section: in the human ratings the BL speakers were also perceived as being more fluent than the IL speakers. On the other hand, these findings are contrary to our expectations and to the results concerning read speech.

However, these findings are less surprising against the background of what was mentioned previously with respect to the speech material used in Experiment 2. In particular, it was suggested that the differences between the items used for the two proficiency groups in Experiment 2 might influence the fluency scores. As explained previously, the short and the long tasks differ not only with respect to length, but also with respect to the nature of the task. More precisely, the BL items contain questions that can be answered immediately by the candidate without much thinking, whereas the IL items contain questions that require more preparation to be answered. In other words, the IL items require more cognitive effort than the BL items, which, in turn, could explain the lower fluency scores, since more cognitively demanding tasks are associated with a lower articulation rate, a lower phonation/time ratio, shorter runs, and longer pauses (Goldman-Eisler, 1968; Grosjean, 1980, pp. 42–43). This is exactly what appears from the comparison of the data for BL and IL in Table IV. The fact that the objective measures also reveal lower fluency in the IL group indicates that the raters in Experiment 2 did a good job and managed to judge fluency independently of proficiency level. The absence of variation in the frequency of filled pauses between proficiency levels might indicate that this phenomenon is not a good indicator of fluency. However, more relevant data in this respect may be provided by the correlation analyses that are presented in the following section.

A final observation about disfluencies concerns the differences between read and spontaneous speech in the specific type of disfluencies produced. To gain more insight into the occurrence of these phenomena in read and spontaneous speech, a manual, more detailed analysis of disfluencies was carried out in which not only quantitative, but also qualitative properties were taken into consideration. Three categories of disfluencies were distinguished: repetitions (exact repetitions of words), repairs (corrections), and restarts (repetitions of initial parts of words). This analysis revealed that the frequency of occurrence of these phenomena varies in the two types of speech. In read speech disfluencies are divided as follows: 12% repetitions, 51% repairs, and 37% restarts. In spontaneous speech, on the other hand, the percentages are: 65% repetitions, 23% repairs, and 12% restarts. These differences between the two distributions appear plausible if one considers that in read speech speakers have to read the words they see on paper and not articulate those which they are planning in their minds. In other words, they are forced, as it were, to pronounce a number of words, some of which might be problematic for them. It is therefore more likely that they will stumble in pronouncing these words, than when they have to pronounce words which they have chosen themselves. It is indeed reasonable to assume that speakers will resort to specific strategies to avoid words that may be difficult to pronounce. However, since they have to formulate the sentences themselves, they will probably need some strategies to win time. This might explain their recourse to repetitions and filled pauses.

TABLE V. Pearson's r correlations between fluency ratings and quantitative measures for read and spontaneous speech. N is indicated in square brackets, while the significance level is indicated by asterisks: *=sign at 0.05 and **=sign at 0.01.

	Read speech	Spontaneous speech	
	all-RS [60]	RBL [28]	RIL [29]
Articulation rate	0.83**	0.07	0.05
Rate of speech	0.92**	0.57**	0.39*
Phonation/time ratio	0.86**	0.46**	0.39*
Mean length of runs	0.85**	0.49**	0.65**
Mean length of silent pauses	−0.53**	−0.08	−0.01
Duration of silent pauses p/m	−0.84**	−0.45**	−0.40*
Number of silent pauses p/m	−0.84**	−0.33*	−0.49**
Number of filled pauses p/m	−0.25	−0.21	−0.21
Number of disfluencies p/m	−0.15	−0.07	−0.27

C. Quantitative measures as indicators of perceived fluency

In this section the automatically calculated temporal measures of speech are compared with the fluency scores assigned by the raters, in order to determine how and to what extent objective measures of speech are related to perceived fluency in read and spontaneous speech. To this end the correlations (Pearson's r) between the two sets of scores in each experiment were calculated. For Experiment 1 the means over the scores assigned by the three rater groups were calculated, because the ratings of the three groups appeared to be very strongly correlated with each other (Cucchiarini *et al.*, 2000). For Experiment 2, on the other hand, the ratings assigned to the two groups of speakers are not directly comparable, because they were assigned by different raters and to different kinds of speech. Consequently, the correlations were calculated for each group of speakers separately. In this way the variation in proficiency level is reduced, which could have consequences for the correlations. All correlations are given in Table V.

The correlations for the read speech material are first considered. As is clear from Table V, all correlations are strong and highly significant (at the 0.01 level), except those for *number of filled pauses per minute* and *number of disfluencies per minute*, which appear not to be statistically significant. Furthermore, it can be observed that although the correlation between perceived fluency and *mean length of silent pauses* is also significant at the 0.01 level, its magnitude is clearly smaller than those of the other coefficients. These results indicate that, at least for read speech, all primary variables are relevant for perceived fluency, the length of silent pauses seems to play a minor role, whereas the secondary variables are not significant. The fact that the number of filled pauses and disfluencies appear not to be good indicators of fluency in read speech is not surprising as these phenomena appeared to be infrequent in read speech. The finding that *mean length of silent pauses* is less related to perceived fluency than the other measures concern-

TABLE VI. Pearson's *r* correlations between fluency ratings and primary variables for read speech at three proficiency levels: PL1 (beginner), PL2 (intermediate), PL3 (advanced). *N* is indicated in square brackets, while the significance level is indicated by asterisks: * = sign at 0.05 and ** = sign at 0.01.

	PL1 [10]	PL2 [27]	PL3 [23]
Articulation rate	0.85**	0.76**	0.66**
Rate of speech	0.92**	0.91**	0.73**
Phonation/time ratio	0.82**	0.86**	0.58**
Mean length of runs	0.91**	0.86**	0.57**
Mean length of silent pauses	−0.50	−0.68**	−0.50**
Duration of silent pauses per minute	−0.71*	−0.85**	−0.61**
Number of silent pauses per minute	−0.83**	−0.83**	−0.57**

ing pauses can be seen as an indication that less fluent speakers, in general, do not make longer pauses than more fluent speakers, but they do pause more often, as was explained in Cucchiarini *et al.* (2000).

When considering the correlations for spontaneous speech, it appears that the secondary variables are not relevant for perceived fluency. In the category of primary variables, on the other hand, there seem to be substantial differences between the various measures. While *rate of speech, phonation/time ratio, mean length of runs, number of silent pauses per minute*, and *duration of silent pauses per minute* exhibit statistically significant correlations with the fluency ratings, *articulation rate* and *mean length of silent pauses* seem to have almost no relation at all with perceived fluency. In particular, *rate of speech* turns out to be the best predictor of fluency for the BL group, while for the IL group *mean length of runs* exhibits the strongest relation with the fluency ratings.

The values in Table V also show that all correlations between the objective measures and the fluency ratings that appear to be statistically significant are systematically lower for spontaneous speech than for read speech. As a possible explanation for this finding one could invoke the lower range in proficiency levels in Experiment 2 as compared to Experiment 1. Recall that in Experiment 1 three proficiency levels were compared and in Experiment 2 only two. In addition, in Experiment 2 this already lower range was further reduced because the correlations were calculated for each proficiency level separately and, therefore, for a rather homogeneous group of speakers as far as proficiency is concerned. To test whether this explanation might be correct, the correlations between the primary variables and the fluency ratings for read speech were computed separately for the three proficiency levels. These correlations are shown in Table VI. As is clear from Table VI, these correlations are slightly lower than those pooled over the three groups, and the correlation concerning *mean length of silent pauses* does not even reach significance in the PL1 group. However, these correlations are still considerably higher than those for spontaneous speech. In other words, even if the range in proficiency is limited by taking a homogeneous group, the correlations for read speech are still higher. Actually, it is amazing that correlations between physical measures and subjective ratings computed on such limited numbers of observations as those in each proficiency level (10 for PL1, 23 for PL2, and 27 for PL3) turn out to be so significantly high. This suggests that the relationship between objective properties and perceived fluency in read speech is rather clearcut, whereas this holds to a lesser extent for spontaneous speech.

However, the most remarkable difference between read and spontaneous speech concerns the correlations between the fluency ratings on the one hand and *mean length of silent pauses* and *articulation rate* on the other. *mean length of silent pauses* already appeared to be less related to perceived fluency than the other variables in read speech, from which it could be concluded that the frequency of the pauses is more important for perceived fluency than their mean length (Cucchiarini *et al.*, 2000). So, in a sense, this finding is less surprising. The absence of a correlation between perceived fluency and *articulation rate*, on the other hand, is much more surprising, because this variable appears to have a strong correlation with perceived fluency in read speech, as is clear from column 2 in Table V. This seems to suggest that when the presence of pauses in speech increases dramatically, as is the case when we go from read speech to spontaneous speech, the contribution of *articulation rate* diminishes and can even become negligible. In other words, even though the dimension of *articulation rate* is perceptually available to the listeners, its effect is overwhelmed by the many pauses.

To determine whether a combination of variables allows one 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. Both for read and spontaneous speech it was found that combining physical parameters does not lead to substantial improvements in predictive power. The results of this analysis show that for read speech the variable that explains the greatest amount of variance is *rate of speech*: R is 0.92; $F = 308.8$; $df = 1$. The second variable that is added in the stepwise procedure is *number of silent pauses per minute*. However, the increase in explained variance is marginal: R rises to 0.93 ($F = 176.5$; $df = 2$). In spontaneous speech the variable that explains the greatest amount of variance is *rate of speech* for the BL group ($R = 0.57$; $F = 12.5$; $df = 1$) and *mean length of runs* for the IL group ($R = 0.65$; $F = 20.2$; $df = 1$). The second variable added in the stepwise procedure is *number of silent pauses per minute* in both cases and in both cases the increase in explained variance is marginal: for the BL group R rises to 0.63 ($F = 8.2$; $df = 2$) while for the IL group R rises to 0.70 ($F = 12.6$; $df = 2$).

IV. DISCUSSION

In this paper the results of two experiments on perceived fluency in read and spontaneous speech have been presented. In these experiments a dual approach was adopted: fluency ratings assigned by experts to read and spontaneous speech produced by non-natives were compared with a number of objective measures that were calculated for the same speech fragments by means of a CSR.

The results of these experiments show that it is possible to obtain reliable ratings of fluency: reliability was high for all rater groups in both experiments (Cronbach's α varied between 0.82 and 0.96). These results may be surprising in view of the much lower degrees of reliability (around 0.68) obtained in previous studies (Riggenbach, 1991; Freed, 1995) and require some explanation. Various factors may have led to such high reliability coefficients in our two experiments in comparison to those in previous studies, in particular, the type of speech analyzed and the raters' degree of experience. First, in Experiment 1 read speech was used while the studies by Riggenbach (1991) and Freed (1995) concerned spontaneous speech. With spontaneous speech, raters have to evaluate fragments that differ not only with respect to fluency in the temporal sense, but also with respect to grammar and vocabulary. As a matter of fact, Riggenbach (1991) and Freed (1995) found that the raters' judgments of fluency were confounded by these linguistic factors. With read speech, on the other hand, these factors can be kept constant so that the raters can concentrate on the temporal variables. In turn, this is likely to result in higher reliability coefficients. Second, a difference in amount of experience in rating speech between our raters and those in Riggenbach's and Freed's experiments may have played a part. The raters in Experiment 2 were professional raters who had received training before starting their activities as raters and had participated in various rating sessions at Cito. The raters in Riggenbach's experiment were ESL instructors and not professional raters. Although ESL instructors are familiar with non-native speech and know how to help learners improve their oral skills, they are probably less used to rating speech in an exam situation than the raters in our experiment. The raters in Freed's experiment were simply native speakers of the language to be rated.

With respect to the major goal of this study, to determine how and to what extent objective properties of speech are related to perceived fluency in read and spontaneous speech, the data analyzed here provide interesting results. First of all, the results obtained in this study have shown how fluency scores, both those assigned by human raters and those obtained on the basis of objective measures, can vary as a function of the type of speech under investigation. Although the human ratings are not readily comparable because they were assigned by different raters, the objective measures do indicate that speakers appear to be less fluent in spontaneous speech than in read speech.

Second, these findings also indicate how the nature of the task carried out by the speaker is related to the fluency scores obtained, both the human ratings and the objective measures. In particular, in presenting the speech material we suggested that the differences between the items used for the two proficiency groups in Experiment 2 might influence the fluency ratings. As explained previously, the short and the long tasks differ not only with respect to length, but also with respect to cognitive load, this being higher for the IL items than for the BL items. In turn, this difference could explain why the speakers in the IL group received lower fluency ratings and had a lower articulation rate, a lower phonation/time ratio, and made longer pauses and shorter runs.

Third, with respect to the role played by the various objective variables these results show that there are both similarities and differences between read and spontaneous speech. The similarities concern the weak relation between the secondary variables and perceived fluency in both types of speech. The differences concern the varying roles of the primary variables in the two speech modalities. As far as read speech is concerned, Table V reveals that the fluency ratings are strongly related to *rate of speech*, *articulation rate*, *phonation/time ratio*, *number of silent pauses per minute*, *duration of silent pauses per minute*, and *mean length of runs*, while *mean length of silent pauses* 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 more fluent and less fluent speakers lies in the number of the pauses they make, rather than in their length, and the longer *duration of silent pauses per minute* of less fluent 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, p. 543) and are corroborated by the data concerning the three proficiency levels: Table IV shows that the differences between the proficiency levels with respect to *mean length of silent pauses* are relatively smaller than those concerning *number of silent pauses per minute* and *duration of silent pauses per minute*. These results suggest that two factors are particularly important for perceived fluency in read speech: the rate at which speakers articulate the sounds and the frequency with which they pause.

With regard to spontaneous speech, Table V shows that the fluency ratings are more strongly related to *rate of speech*, *phonation/time ratio*, *number of silent pauses per minute*, *duration of silent pauses per minute*, and *mean length of runs*, while *articulation rate* and *mean length of silent pauses* have almost no relationship with perceived fluency. Since pauses are much more frequent in spontaneous speech than in read speech (see Table IV), it is possible that their prominence effaces the importance of *articulation rate*. In other words, in speech where pauses are very frequent it seems plausible that a variable that takes no account of pauses whatsoever, like *articulation rate*, shows no relation to perceived fluency. Furthermore, when one considers the nature of all these variables it appears that fluency ratings of spontaneous speech are particularly related to variables that contain information about the frequency of the pauses, and these are *rate of speech*, *phonation/time ratio*, *number of silent pauses per minute*, *duration of silent pauses per minute*, and *mean length of runs*, but not *articulation rate* and *mean length of silent pauses*. In turn, this suggests that of the two factors that are important for perceived fluency in read speech, namely the rate at which speakers articulate the sounds and the frequency with which they pause, the latter is most important for perceived fluency in spontaneous speech.

Another interesting finding in this study is that *mean length of runs* is a particularly good predictor of fluency in spontaneous speech and for the IL group it is better than all other measures that do take pause frequency into account. What distinguishes *mean length of runs* from the other measures is that *mean length of runs* takes account not only of the frequency of the pauses but, to a certain extent, also of their distribution. The importance of this variable seems to suggest that pauses are tolerated, provided that sufficiently long uninterrupted stretches of speech are produced. The fact that the predictive power of *mean length of runs* is greater for the IL group, i.e., for speech material where the speaker has to present his/her arguments in a coherent and organized manner and where the distribution of pauses is of course more important, lends further support to this interpretation.

In our previous paper (Cucchiarini *et al.*, 2000) it was noted that a possible limitation of that study was that it only indicated a strong relationship between objective measures of temporal speech characteristics on the one hand and expert fluency ratings on the other, but it did not provide information as to how varying articulation rate and/or pause time would affect the fluency ratings. In other words, we admitted that we were not in a position to make strong claims about the causal relationships obtaining between the objective measures and the fluency ratings. By using speech where a different relationship between articulation rate and pause time obtains, such as the spontaneous speech used in the present study, we have attempted to get more insight into how variations in the objective properties of speech affect fluency ratings. As a matter of fact, it turned out that as pauses become more frequent, as in spontaneous speech, the importance of articulation rate is reduced.

At this point it is interesting to consider what implications the results of this study can have for the future of fluency assessment. In our previous paper (Cucchiarini *et al.*, 2000) we expressed our optimism with respect to the potential that our approach could have for objective fluency assessment in read speech by identifying quantitative speech variables that appear to be strongly related to perceived fluency. The extension to spontaneous speech analyzed in the present paper has shown that the predictive power of the various quantitative variables may differ for read and spontaneous speech and that, even within the same speech type, fluency ratings may vary depending on the specific task carried out by the speaker. In this respect the results of this study confirm those of previous investigations which indicated that cognitively more demanding tasks lead to lower fluency scores. For fluency assessment these findings imply that objective measures can be employed, but that the specific selection of variables and their interpretation should be related to the type of speech and the type of task.

V. CONCLUSIONS

On the basis of the results of the present study the following conclusions can be drawn. First, expert listeners are able to evaluate fluency with a high degree of reliability, both in read and in spontaneous speech. Second, fluency scores, both those assigned by human raters and those calculated on the basis of objective properties of speech, appear to vary with the type of speech under investigation: speakers turn out to be more fluent in read than in spontaneous speech. Third, fluency scores also vary with the type of task carried out by the speaker, with cognitively more demanding tasks being associated with lower fluency scores. Fourth, while expert fluency ratings of read speech are mainly related to speed of articulation and frequency of pauses, those of spontaneous speech appear to be more related to the frequency and distribution of pauses while speed of articulation shows almost no relation to perceived fluency. Fifth, expert fluency ratings can be more accurately predicted in read speech than in spontaneous speech 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 in almost all cases. The only exceptions are the cognitively more demanding tasks in the spontaneous speech experiment for which *mean length of runs* turns out to be the best predictor of fluency.

To conclude, these findings indicate that temporal measures of fluency may be employed to develop objective testing instruments of fluency in read and spontaneous speech. However, the selection of the variables to be employed in such tests should be dependent on the specific type of speech material investigated and the specific task performed by the speaker.

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. We are indebted to 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.

Bortfeld, H., Leon, S. D., Bloom, J. E., Schober, M. F., and Brennan, S. E. (**1999**). ''Which speakers are most disfluent in conversation and when?,'' Proceedings ICPhS99 Satellite Meeting on Disfluency in Spontaneous Speech, pp. 7–10.

Boves, L. (**1984**). *The Phonetic Basis of Perceptual Ratings of Running Speech* (Foris, Dordrecht).

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

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

Cucchiarini, C., Strik, H., and Boves, L. (**2000**). ''Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology,'' J. Acoust. Soc. Am. **107**, 989–999.

Den Os, E. A., Boogaart, T. I., Boves, L., and Klabbers, E. (**1995**). ''The Dutch polyphone corpus,'' Proceedings Eurospeech95, pp. 825–828.

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 (Benjamin, 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 Francais: Vitesse de parole et variables composantes, phénomènes d'hésitation,'' Phonetica **31**, 144–184.

Laver, J. (**1980**). *The Phonetic Description of Voice Quality* (Cambridge University Press, Cambridge).

Leeson, R. (**1975**). *Fluency and Language Teaching* (Longman, London).

Lennon, P. (**1990**). ''Investigating fluency in EFL: A quantitative approach,'' Language Learning **3**, 387–417.

Nation, P. (**1989**). ''Improving speaking fluency,'' System **3**, 377–384.

Neumeyer, L., Franco, H., Weintraub, M. and Price, P. (**1996**). ''Automatic text-independent pronunciation scoring of foreign language student speech,'' Proceedings of the International Conference on Spoken Language Processing (ICSLP) '96, pp. 1457–1460.

Raupach (**1983**).

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.

SPEX http://lands.let.kun.nl/spex.

Strik, H., Russel, A., Van den Heuvel, H., Cucchiarini, C., and Boves, L. (**1997**). ''A spoken dialog system for the Dutch Public Transport Information Service,'' Int. J. Speech Technol. **2**, 121–131.

Towell, R., Hawkins, R., and Bazergui, N. (**1996**). ''The development of fluency in advanced learners of French,'' Appl. Linguist. **1**, 84–119.

Townshend, B., Bernstein, J., Todic, O. and Warren, E. (**1998**). ''Estimation of spoken language proficiency,'' Proceedings of the ESCA Workshop Speech Technology in Language Learning (STiLL 98), pp. 179–182.

Van Bezooijen, R. (**1984**). *Characteristics and Recognizability of Vocal Expressions of Emotions* (Foris, Dordrecht).