Quantitative Correlates as Predictors of Judged Fluency in Consecutive Interpreting: Implications for Automatic Assessment and Pedagogy

Wenting Yu[1] and Vincent J. van Heuven[2, 3]
[1] Shanghai International Studies University / [2] University of Pannonia, [3] Leiden University

Abstract

This chapter presents an experimental study of consecutive interpreting which investigates whether: (1) judged fluency can be predicted from computer-based quantitative prosodic measures including temporal measures and melodic measures. Ten raters judged six criteria of accuracy and fluency in two consecutive interpretations of the same recorded source speech, from Chinese ‘A’ into English ‘B’, by 12 trainee interpreters (seven undergraduates, five MA students). The recorded interpretations were examined with the speech analysis tool PRAAT. From a computerized count of the pauses thus detected, together with disfluencies identified by raters, 12 temporal measures of fluency were calculated. In addition, two melodic measures, i.e. pitch level and pitch range were automatically generated. These two measures are often considered to be associated with speaking confidence and competence. Statistical analysis shows: (1) strong correlations between judged fluency and temporal variables of fluency; (2) no correlation between pitch range and judged fluency, but a moderate (negative) correlation between pitch level and judged fluency; and (3) the usefulness of effective speech rate (number of syllables, excluding disfluencies, divided by total duration of speech production and pauses) as a predictor of judged fluency. Other important determinants of judged fluency were the number of filled pauses, articulation rate, and mean length of pause. Potential for developing automatic fluency assessment in consecutive interpreting is discussed, as are implications for informing the design of rubrics of fluency assessment and facilitating formative assessment in interpreting education.

Keywords: fluency, quantitative correlates, consecutive interpreting, automatic assessment

1. Introduction

Fluency is an important aspect of quality assessment of natural speech and also of interpreting. Though fluency alone provides no guarantee of the interpreter’s reliability unless content too is taken into account, it is widely recognized as a feature of successful interpretation (Mead 2005). Since the 1980s, fluency has been studied as one out of many aspects of quality in interpreting and its role in interpreting quality assessment has gained an increasing amount of recognition. It is nevertheless only recently that interpreting scholars have begun to foreground fluency as a specific component of quality, with the aim of defining it in practical terms and determining to what extent it might affect intelligibility and user perception (Rennert 2010). So far, a number of empirical studies have indicated that: (a) there is a contrast between user expectations and user perception regarding the importance of fluency as a quality criterion in interpreting; and (b) fluency is such an important element in perceived quality as its experimental manipulation affects user perception of other criteria (Collados Aís, Pradas Macías, Stévaux & García Becerra 2007; Pradas Macías 2003, 2007; Rennert 2010). These studies argue that the importance of fluency had long been underestimated, given that previous surveys of quality perception or expectations among interpreters and habitual listeners showed a tendency to prioritize accuracy or faithfulness over form-related features like fluency (Bühler 1986; Chiaro & Nocella 2004; Kurz 1993, 2001, 2003; Moser 1996).

Fluency is a polysemous and somewhat elusive concept, though it is frequently applied to describe oral language performance. Since there has been no generally agreed definition of fluency and since it has been used to characterize a wide range of different skills and speech characteristics such as the temporal aspects of speech, the degree of coherence and semantic density in speech, and the appropriacy of speech content in different contexts (Leeson 1975; Fillmore 1979; Brumfit 1984; Lennon 1990; Schmidt 1992; Chambers 1997), assessment of fluency in interpreting is considered a difficult task. In everyday language use, fluency is often considered as a synonym of general language proficiency (Lennon 1990; Chambers 1997). Accordingly, fluency is often associated with overall quality of interpreting. In light of a more restricted definition of fluency that lays emphasis on “native-like rapidity” (Lennon 1990, p.390), assessors of interpreting may be more concerned with how an interpretation is delivered and the way an interpreter controls the use of silent pauses, hesitations, filled pauses, self-corrections, repetitions, false starts and the like. Thus, the various definitions of fluency may largely contribute to the variability in assessment standards and lack of consensus in fluency assessment. Another difficulty of fluency assessment in interpreting is that assessors’ judgement is very likely to be subject to the “impressions” after listening to the interpretation, because an interpreter’s work is ephemeral in nature and even the same assessor might not be consistent in assessment at different times. In this case, it is also very likely that assessors may find it difficult to dissociate fluency with other quality measures when their “impressions” of certain indices (e.g., fidelity, information completeness) become dominant.

In response to the existing problems described above, Mead (2005) proposed that objective and quantitative measurement of temporal measures of fluency could be more transparent and reliable than assessment based on content-related parameters such as completeness and correctness, which is often subject to differing opinions of what is right, wrong or missing and is ultimately more difficult to pin down (Mead 2005). Admittedly, it is important to recognise that fluency as a key, form-related feature of successful interpretation may not reveal the quality of interpreting without taking content into account. However, a quantitative perspective on different features of interpreting such as fluency can contribute to overall assessment of quality (Mead 2005). Though the quantitative approach to assessment has only recently been introduced to interpreting studies, it has long been practiced in more established disciplines like language testing and educational assessment. Studies on L1 and L2 speech have attempted to define fluency in terms of objective speech properties (Cucchiarini, Strik & Boves 2000, 2002; Freed 1995; Lennon 1990; Riggenbach 1991). According to Cucchiarini et al. (2002), these studies have adopted a dual approach in which fluency is evaluated by listeners and quantified by temporal measures, and in which subjective evaluations are usually related to objective measures. This type of approach, particularly useful for gaining insight into the acoustic features underlying listeners’ evaluations, has a long tradition in phonetic research (Cucchiarini et al. 2002). It has been observed that, for speech tasks entailing different degrees of cognitive effort, there will be corresponding differences in fluency rating (Bortfeld, Leon, Bloom, Schober & Brennan 1999; Grosjean 1980) and in objective and acoustic predictors of fluency (Cucchiarini et al. 2002). It is therefore reasonable to expect that, for the cognitively complex task of interpreting, relations between perceived fluency and quantitative measures will differ from those in everyday L1 and L2 speech. In other words, the quantitative measures that are highly predictive of judged oral fluency of unconstrained L1 and L2 production might not retain their predictive power for interpreted rendition.

In Yu and van Heuven (2017), our study applied the dual approach to gain insights into the temporal aspects of fluency as indicators of perceived fluency in consecutive interpreting (CI) (from Chinese 'A' into English 'B'). That study was innovative in two important aspects. First, it extended the existing research on the effect of task complexity on the linguistic output of fluency. Second, it investigated the relation between temporal parameters of fluency and judged fluency in CI, probing into which temporal measure(s) of fluency were most indicative of fluency in CI.

In this chapter, based on the same data set generated from our previous experiment (Yu & van Heuven 2017), we further expand our study by extracting new data of two quantitative prosodic measures, namely median vocal pitch and size of effective pitch range, which we hypothesized to predict fluency in CI like the temporal parameters we had identified (Yu & van Heuven 2017).

As explained previously, the importance of this kind of research is two-fold. It not only generates a better understanding of how perceived fluency would relate to quantitative characteristics of delivery in CI, but also produces implications for the future development of quantitative testing instruments. Ultimately, such research may contribute to greater testing efficiency and better learning results in interpreter training.

2. Quantitative Assessment of Fluency

It is a common practice that fluency is used to describe oral language performance, and sometimes written performance (e.g. Lennon 1990). A review of relevant literature reveals that fluency has been used to characterize different skills and speech properties, but there has been no consensus on its definition in different contexts (e.g., Brumfit 1984; Chambers 1997; Fillmore 1979; Leeson 1975; Lennon 1990, 2000; Schmidt 1992). Cucchiarini et al. (2000) distinguishes native language fluency from fluency in the context of foreign language teaching and testing. Native language fluency is used to characterize the performance of a speaker, and is often considered a synonym of "overall language proficiency" (Chambers 1997; Lennon 1990), but it does not really constitute an evaluation criterion (Cucchiarini et al. 2000). By contrast, fluency in foreign language teaching and testing tends to be formally used as a criterion of evaluation by which non-native performance can be judged (Riggenbach 1991; Freed 1995), despite the variability and vagueness of its definitions. In a more restricted sense, fluency can refer to temporal aspects of oral proficiency (Freed 1995; Lennon 1990; Nation 1989; Riggenbach 1991; Schmidt 1992; Towell, Hawkins & Bazergui 1996), in line with Lennon's (1990) assumption that the goal in foreign language learning consists in producing "speech at the tempo of native speakers, unimpeded by silent pauses and hesitations, filled pauses, […]

self-corrections, repetitions, false starts and the like.”

The identification of temporal features of speech is the prerequisite for quantitative studies of fluency in different contexts: previous research shows that perceived fluency can be correlated with different quantitative measures, depending on the language and specific speech task – e.g., L1 speech vs. L2 speech; read speech vs. spontaneous speech (Cucchiarini et al. 2002; Kormos & Dénes 2004; Möhle 1984; Towell et al. 1996).

The focus of the present study is to investigate the relationship between judged fluency and quantitative measures of fluency in the cognitively demanding speech task of CI from Chinese ‘A’ into English ‘B’ (or from L1 to L2). For our trainee interpreters who are native Chinese speakers, there is always a major gap between A (Chinese) and B (English) language proficiency. Accordingly, oral proficiency in B language often receives more attention in the interpreting syllabus. Therefore, it is necessary to review the existing studies on fluency in two types of L2 speech tasks (i.e., read speech and spontaneous speech), in which listeners' evaluations of speech are examined in relation to temporal and prosodic measures calculated for the same speech samples.

Research on fluency in L2 speech mostly aims to gain insight into the factors that underpin listeners' evaluations (Cucchiarini et al. 2002), and/or to help develop objective tests of L2 fluency that might lead to automatic assessment (Cucchiarini et al. 2002; Townshend, Bernstein, Todic & Warren1998). A number of researchers, including Lennon (1990), Riggenbach (1991), Freed (1995), Kormos and Dénes (2004), Cucchiarini et al. (2002) and Pinget, Bosker, Quené, Sanders and de Jong (2014), 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 (the percentage of speaking time used for actual speech production), mean length of run, and number and length of pauses. In Cucchiarini et al.'s (2002) study, the relationship between objective fluency measures of speech and perceived fluency in L2 Dutch read and spontaneous speech was investigated in two separate experiments. They found that fluency ratings in both cases were closely related to speech rate, phonation/time ratio, number of silent pauses per minute, duration of silent pauses per minute, and mean length of run. While articulation rate showed almost no relationship with the perceived fluency ratings in spontaneous L2 speech, the two were closely correlated in read L2 speech production. The authors' tentative explanation for this finding was that, since pauses tended to occur much more frequently in spontaneous speech, articulation rate (which takes no account of pauses) may in practice be relegated to a position of irrelevance. Kormos and Dénes (2004) conducted a study on L2 Hungarian spontaneous speech fluency ratings and temporal measurements, and reported that speech rate, mean length of utterance, phonation/time ratio and the number of stressed words produced per time unit were the best predictors of fluency scores. Like Cucchiarini et al. (2002), they did not find that articulation rate, the number of filled and unfilled pauses, or other disfluency phenomena were good predictors of fluency ratings. A more recent study by Pinget et al. (2014) investigated which acoustic measures of fluency can predict perceived fluency in L2 Dutch spontaneous speech. Although their acoustic measures (calculated on the basis of syllable length and pause length/frequency) differed from those used in previous research, these parameters showed high predictive power for much of the variance in fluency ratings, while two measures of repair fluency (number of corrections and number of repetitions) showed a certain – albeit limited – degree of predictability compared to other studies (cf. Cucchiarini et al. 2002; Kormos & Dénes 2004). These studies which scrutinized different L2 languages show that: (1) fluency ratings are mainly affected by temporal variables related to speed fluency (i.e., the speed at which speech is delivered) and breakdown fluency (i.e., the number and length of pauses); (2) the relationship between fluency ratings and temporal variables in spontaneous speech may be rather complex, since the former can be affected by non-temporal language features such as grammar, vocabulary and accent (Freed 1995; Lennon 1990; Riggenbach 1991).

To sum up, a number of studies have shown that quantitative assessment can be used to identify objective measures that are predictive of subjective fluency ratings in L2 speech. The general consensus is that temporal measures related to speed fluency and breakdown fluency are far more predictive than repair fluency. However, the predictive power of objective measures differs for different L2 speech tasks depending on the cognitive effort involved.

3. Applying Quantitative Assessment of Fluency to CI

Presumably, interpreting is more complex than L2 oral tasks, regarding the cognitive processes involved. What distinguishes CI from everyday spoken language activity is readily appreciated by basing comparison of the two on models often used to analyse them: Levelt's (1989) speech production model and Gile's (1995) Effort Models. The main difference between the two is that the speech production model has an initial conceptualization stage, whereas CI starts with perception and comprehension of the source language, with parallel storage, processing, and retrieval of information through note-taking, memory functions, and coordination of all these efforts. As a result, more attentional resources are almost certainly required in CI than in spontaneous speech production.

Assessment of interpreting quality is necessary in both professional practice and interpreter training, which is generally based on two broad aspects: (a) content-related features such as accuracy and completeness of information; and (b) form-related features represented by fluency of delivery, accent, intonation, and voice quality (e.g., Bühler 1986; Zwischenberger & Pöchhacker 2010). Fluency is among the most important formal criteria, and it contributes to the overall quality of interpreting, which is testified by the fact that it forms part of numerous rating scales in interpreting tests (Han et al., 2020). However, it seems to have attracted little attention in the teaching and training of interpreting. Recent research has examined how users' expectations could differ from their actual assessments of interpreting fluency. Research results suggest that limited fluency may impact negatively on the overall judged quality of an interpretation (e.g., Collados Aís 1998; Pradas Macías 2003; Rennert 2010). In a large-scale global survey on conference interpreting quality involving 704 AIIC interpreters worldwide (Pöchhacker 2012), fluency was perceived as being very important by 71% of participants and ranked third out of eleven quality criteria (behind sense consistency and logical cohesion). In Yu and van Heuven's (2017) study, judged fluency in CI was significantly correlated with judged accuracy. These studies indicate that the importance of fluency as one of the key quality indicators of interpreting performance is attracting a growing amount of attention. Against this backdrop, research on quantitative approaches to fluency in interpreting seems both indispensable and practical. It is indispensable because among the problems related to assessment method, judging-by-impression by assessors is most likely to result from the evanescent nature of interpreting, rendering the assessment to be subjectivity-based. Even for experienced assessors, lack of consistency between the various assessments might occur, which indicates considerable variability in standards and priorities from one assessor to another (Mead 2005). It is practical because a number of acoustic parameters that underpin fluency (e.g., speech rate, the number and length of un/filled pauses, the number of false starts and self-repairs) make it possible to examine fluency through objective measurement.

Compared with an abundant amount of empirical studies on quantitative assessment of fluency in L2 speech (see Section 2), relatively little research has been conducted to explore quantitative temporal parameters that potentially underlie fluency with a few notable exceptions (Mead 2005; Yang 2015; Yu & van Heuven 2017; Han et al. 2020). Based on the analysis of five temporal parameters of fluency in Mead's (2005) pioneering work on elaborating a conceptual approach to quantitative assessment of fluency in interpreting, he suggested that speech rate, pause duration, and length of fluent run can be taken as the most relevant parameters in assessing interpreting fluency. Yang's (2015) exploratory study attempted at relating temporal measures of utterance fluency to perceived fluency ratings in an exploratory study participated by 18 postgraduate student interpreters consecutively interpreting from Chinese to English. Her results indicated that overall speech rate (syllables per minute), articulation rate (pruned, syllables per minute), phonation time ratio, and mean length of silent pauses (pause cut-offs set at 0.3, 0.4 and 0.5 seconds) were closely related to the subjective fluency ratings. In a similar experimental study involving 12 trainee interpreters (7 undergraduate students and 5 postgraduate students) by Yu and van Heuven (2017) on the correlations between judged fluency and temporal predictors of fluency in CI (Chinese to English), they included in their study repair disfluencies such as false starts, restarts, corrections and repetitions and identified effective speech rate (number of syllables, excluding disfluencies, divided by total duration of speech production and pauses) as the most predictive temporal parameter among all together 12 temporal fluency measures. Most recently, Han et al. (2020) made an extended study with 41 undergraduate interpreting students as participants, aiming at modeling the relationship between utterance fluency and raters' perceived fluency of CI. They reported that mean length of unfilled pauses, phonation time ratio, mean length of run and speech rate had fairly strong correlations with perceived fluency ratings in both interpreting directions (English to Chinese and Chinese to English) and across rater types.

In addition, melodic features such as intonation and accent may potentially play a role in influencing perceived fluency. Melodic features, together with temporal features constitute two broad categories of phenomena of prosody (van Heuven 1994, 2017). Like temporal features, melodic features are also part of quality criteria in interpreting which usually are deemed less important than content-related criteria such as accuracy and logical cohesion (Collados Aís et al. 2007; Pöchhacker 2012). However, empirical studies have shown that melodic features in interpreting delivery seem to influence 1) users' understanding of the speech content (Holub & Rennert 2011; Shlesinger 1994; Déjean 1990); 2) their perception of overall interpreting quality (Cheung 2013); and 3) the perception of fluency (Christodoulides & Lenglet 2014). Most interestingly, experimental research has found that listeners asked to rate "fluency" seem to confluence temporal features with intonation (e.g. pitch variation) (Collados Aís et al. 2007). Therefore, it seems appropriate to incorporate quantifiable melodic parameters in the present study.

This experimental study aims to address two related research questions. The first research question is: What are the correlations between raters' subjective assessments of consecutive interpreters' fluency and quantitative measurements of both temporal and melodic parameters? This is followed by the second research question: Which quantitative prosodic measure(s) can best predict judged fluency in CI?

4. Methods

4.1 *Participants*

Twelve students from Shanghai International Studies University participated in this study: seven third-year BA translation and interpretation majors, with a mean age of 20, and five second-year MA students, with a mean age of 25, from the Graduate Institute of Interpretation and Translation. The BA students were still working on development of basic interpreting skills, while the MA students were already interpreting part-time and were working towards their professional qualification. By the time of the experiment, the BA students had completed three basic one-semester CI training courses; the MA students had undergone three semesters of intensive and advanced interpreter training (at least three hours a day, covering both CI and SI). All participants had Chinese A and English B.

4.2 *Material*

A source audio clip in Chinese (3.5 minutes in duration, with a total of 501 Chinese characters comprising six paragraphs) was prepared from recordings of the press conference (2.5 hours) held by the former Chinese Premier Wen Jiabao during the National People's Congress in 2009. The audio clip was played to the student participants and they interpreted it consecutively into English. Of the six interpreted paragraphs, two (paragraphs 4 and 5) were selected for perceptual rating and acoustic analysis. The reason for focusing on an extended extract, rather than the whole interpretation, was that the rating task had to be manageable so as to avoid rater fatigue.

4.3 *Procedures*

4.3.1 *Experiment*

The experiment was originally designed and run to test an earlier hypothesis regarding improvement of judged fluency when exactly the same speech is interpreted a number of times in quick succession (Yu & van Heuven 2013). The experiment took place in conference rooms equipped with booths for simultaneous interpreting. The source stimulus material and the participants' rendition were digitally recorded on separate tracks to maintain time differences. One of the authors monitored the interpreters over headphones and ensured that all of them had finished interpreting one paragraph before the next was played to them. The participants were instructed to interpret the same source speech three times (deliveries 1, 2 and 3), paragraph by paragraph, with a break of two minutes between deliveries. Delivery 1 and delivery 3 were selected for both auditory rating and acoustic analysis. Delivery 2 was excluded, because previous studies (e.g., Zhou 2006) suggest that the third delivery is often the most proficient during oral task repetition.

4.3.2 *Fluency ratings*

The online survey software *Qualtrics* was used for the rating procedure. Twenty-four clips (12 interpreters × 2 interpretations each) were rated on six measures related to accuracy and fluency: (i) accuracy of information; (ii) grammatical correctness; (iii) speed of delivery; (iv) control of pauses (both silent and filled); (v) control of other disfluencies (unnecessary repetitions, false starts, inappropriate lengthening of syllables, self-corrections); and (vi) overall fluency, on a scale of 1- 10. The presentation of the 24 clips was randomized for each rater, to prevent the potential order effect.

Ten raters (five men, five women), with a background of studying or teaching at Leiden University, participated in the online rating: three native English speakers (two UK, one US), and seven with near-native English proficiency (six Dutch L1 and one Portuguese L1, all members of the academic staff in the English section, or PhD candidates in linguistics). The raters were informed that the entire rating session would last an hour and advised to take a ten-minute break after rating twelve clips, so as to avoid fatigue. They were then asked to complete a background questionnaire, after which they carefully read through an English translation of the two paragraphs so that they understood the messages to be interpreted. Subsequently, the audio clips of two specimen interpretations were played to them: one very good, the other less so. These were recordings of interpreters who were not actually included in the experimental sample.

The ten raters scored all the twelve participants across two deliveries, as explained above. Means were calculated, to obtain one score for each single delivery on each rating measure. A total of 24 ratings was thus obtained for each of the six rating measures.

4.3.3 *Acoustic correlates of fluency*

The following measures were selected for investigation:

(1) articulation rate = number of syllables, including disfluencies, divided by total duration of speech apart from all (silent and filled) pauses longer than 0.25 seconds;
(2) speech rate = as for articulation rate, but including all pauses in the total speech duration;
(3) effective speech rate = as for speech rate, but excluding disfluencies from the syllable count;
(4) number of silent pauses above 0.25 seconds in duration;
(5) mean length of silent pauses longer than 0. 25 seconds;
(6) number of filled pauses (*uh*, *er*, *mm*, etc.);
(7) mean length of ~~all~~ filled pauses;
(8) number of pauses = sum of (4) and (6).
(9) mean length of pauses = mean of (5) and (7), weighted by their respective frequencies (items 4 and 6);
(10) number of other disfluencies (repetitions, restarts, false starts, corrections);
(11) mean length of fluent runs = mean number of syllables produced between silent pauses

longer than 0.25 seconds;

(12) phonation/time ratio, calculated on the basis of items 4 and 6, as a percentage of overall speech time = (total duration of speech without pauses, divided by total duration of speech including pauses) × 100

(13) median pitch = the median value of the fundamental frequency (in hertz, Hz)

(14) size of effective pitch range = difference between the 10th and 90th percentile of the pitch distribution (in semitones)

The threshold of 0.25 seconds for silent pauses is used to distinguish hesitation in speech (Towell, Hawkins & Bazergui 1996) from pauses that are part of normal articulation for some combinations of sounds or may be classified as micro-pauses (Pinget et al. 2014; Riggenbach 1991).

According to Tavakoli and Skehan's (2005) perspective on the temporal properties of fluency, the above acoustic measures are predictive of speed fluency (1, 2), breakdown fluency (2, 4, 5, 6, 7, 8, 9, 11 and 12), repair fluency (10), or all three categories (3).

In the present study, speech length is measured in syllables. This is in line with Pöchhacker's (1993) observation that use of syllables as a standard international unit of measurement obviates the practical drawback caused by the sometimes considerable variability in word length across different languages. Calculation of some temporal measures (e.g., phonation/time ratio) requires a count of the transcribed syllables: in our study, this was done manually by the first author and checked by a student assistant.

For all 24 clips, the transcription was made by a graduate student assistant and checked by the first author. The transcriber was instructed to listen very carefully, noting any apparently unpronounced syllables for the purpose of the subsequent syllable count. No syllables had actually been omitted. Detailed phonetic transcription and a manual syllable count would be laborious for longer speech samples. Automatic syllable count can be realized by running a script in PRAAT developed by de Jong and Wempe (2009), with the prospect of further improvements for future use in research such as this. Automatic phonetic transcription is also very likely to be facilitated, as automatic speech recognition technology evolves. Currently, however, we consider that manual calculation offers the best guarantee of accuracy.

The transcription of the 24 clips included filled pauses and all types of disfluencies. Silent pauses were detected by running the MarkInterval script (developed by Jos Pacilly) in PRAAT~~:~~.[1] ~~this~~ The software ~~made it possible to~~ converts an acoustic signal into an oscillogram and/or spectrogram, visualizing sounds as a continuous wave pattern in which any segment can be matched with the corresponding recording. At a sampling frequency of 44.1 KHz, duration of different speech features can be measured in milliseconds. Together with the oscillogram, two annotation levels ('tiers' in a 'textgrid') were created for the transcribed texts and the labelled disfluencies (see Figure 1 for an example). The length of each silent and filled pause detected, the total duration and number of all pauses, and the number of disfluencies were automatically extracted.

Silences at the very beginning and end of every delivery were discarded. The selection of the variables in this study is slightly different from Cucchiarini et al.'s (2002) choice of nine

[1] The MarkInterval script was written by Jos J.A. Pacilly, speech software specialist at the Leiden University Centre for Linguistics.

temporal variables related to fluency in L2 spontaneous speech. First, a distinction was made between effective speech rate (a variable proposed by us) and speech rate in general. Effective speech rate is calculated after excluding syllables identified as disfluencies (e.g., involuntary repetitions), because these are very likely to be more frequent in the cognitively demanding speech task of interpreting than in unconstrained speech. Second, syllables were used as the units of measurement. Third, mean length of filled pauses was added. The rationale for this decision was that interpreting is probably more conducive than spontaneous speech to hesitation pauses, as the interpreter takes time to analyze incoming information while also planning and retrieving the components of target language production. Finally, the number and mean length of pauses were also added so as to measure overall pausing.

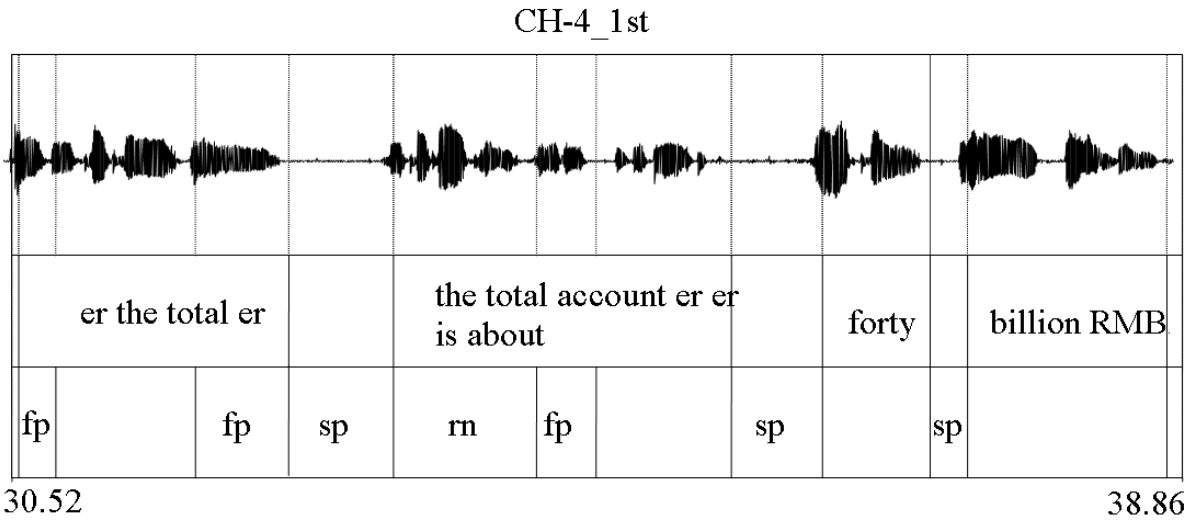

Figure 1. Visualization of a selected sound wave, with two annotation tiers in a text grid, in PRAAT (CH4_1st = Delivery 1 by the participant interpreting in Booth No. 4; fp = filled pause; sp = silent pause; rn = repetition).

Before we set out to generate the pitch information for each of the 24 audio clips using PRAAT, we set an analysis range (the upper boundary and the lower boundary of pitch) between 125 and 400Hz for most participants with the exception of two female and one male. As the clips of one of the two female participants had patches containing creaky voice (extremely low pitch), we used a routine procedure to unvoice the creaky parts and adjusted the analysis range to 150-350Hz. The clips of the other female participant had a rather low-pitched voice for a woman, so we set the pitch range between 75 and 250Hz, which is the same analysis range we applied to the clips by the male participant. Besides, we also unvoiced the parts where some participants produced jump-ups of pitch when pronouncing affricate and fricative sounds such as /tʃ/ and /ʃ/. The effective size of pitch range was calculated in semitones as the difference between the 90th and the 10th percentiles of the speaker's pitch distribution. The median pitch was the exact value in hertz (Hz) of the 50th percentile of the pitch distribution.

The quantitative measures described above were calculated separately for every single interpretation (i.e., two values per participant for each parameter). A total of 24 values was thus obtained for each quantitative prosodic measure. Correlations between these values and the judged fluency ratings were then analyzed.

4.4 *Statistical analysis*

The ten raters were highly consistent in their ratings on the six measures related to accuracy and fluency, with a mean Cronbach's *α* of .96. The highest *α* was .97, for grammatical correctness and speed of delivery; the lowest *α* was .94, for accuracy of information. Table 1 shows the Cronbach's alpha as a measure of internal consistency for the six fluency rating items.

Table 1. Internal consistency (Cronbach's *α*) of the ten raters for each rating item~~.~~

Rating item	***α***
Accuracy of information	.94
Grammatical correctness	.97
Speed of delivery	.97
Control of pauses	.95
Control of disfluencies	.96
Overall fluency	.95

For the statistical analyses of the rating results and quantitative measures, Pearson's *r* and (multiple) linear regression were performed.

5. Results

The results of the accuracy and fluency ratings assigned by the ten raters are presented first, followed by those for the quantitative acoustic measures of fluency. Finally, the relationship between the two sets of measures is studied.

5.1 *Auditory ratings of fluency and accuracy*

We present the mean scores on the six ratings for accuracy and fluency parameters (see Table 2). The ratings differed for the beginners and the advanced students, as well as for deliveries 1 and 3. The analysis in our earlier study (Yu & van Heuven 2013) showed significant main effects of both repetition (delivery) and proficiency on perceived accuracy and fluency. This means that the advanced students were judged as being significantly more accurate and fluent than the beginners, while all students were judged as being significantly more accurate and fluent in delivery 3 than in delivery 1. In addition, the ratings for accuracy and fluency-related measures were found to correlate strongly in Yu and van Heuven's (2017) study (see Table 3).

Table 2. Mean ratings for the six measures related to accuracy and fluency.

		Delivery 1						Delivery 3					
		Accuracy		Fluency				Accuracy		Fluency			
Proficiency	ID	AI	GC	SD	CP	CD	OF	AI	GC	SD	CP	CD	OF
BA	1.	5.7	5.7	5.7	4.8	4.8	5.3	6.1	6.4	6.7	6.1	5.4	6.1
BA	2.	5.6	5.2	5.5	4.9	5.0	5.6	7.5	6.6	6.5	5.5	5.6	6.3
BA	3.	4.5	4.6	5.0	4.7	4.3	4.5	6.4	6.0	6.2	5.9	5.9	6.2
BA	4.	5.2	4.9	4.4	4.0	3.6	4.1	5.1	5.2	5.9	5.6	5.2	5.7
BA	5.	5.6	5.3	5.2	5.0	4.8	5.1	6.8	6.6	6.1	6.4	6.0	6.2
BA	6.	4.7	5.2	4.9	4.6	4.4	4.8	6.7	5.9	5.5	5.6	6.0	5.9
BA	7.	5.5	5.7	5.0	3.9	4.4	4.4	5.7	6.0	6.3	5.7	5.1	5.7
Mean (BA)		5.3	5.2	5.1	4.6	4.5	4.8	6.3	6.1	6.2	5.8	5.6	6.0
MA	8.	7.0	6.6	7.0	6.4	6.1	6.8	7.5	6.5	7.3	6.9	6.5	7.2
MA	9.	7.6	6.8	7.7	6.7	6.4	7.1	8.1	7.6	7.9	7.6	7.4	7.6
MA	10.	7.8	6.9	7.6	6.9	6.8	7.3	8.2	7.7	8.0	7.2	7.2	7.8
MA	11.	7.1	5.7	6.3	5.7	5.7	6.1	7.4	6.0	6.8	6.6	6.2	6.6
MA	12.	5.6	5.5	5.6	4.8	4.7	5.3	6.2	6.7	6.7	5.8	5.2	5.8
Mean (MA)		7.0	6.3	6.8	6.1	5.9	6.5	7.5	6.9	7.3	6.8	6.5	7.0
Mean (grand)		6.0	5.7	5.8	5.2	5.1	5.5	6.8	6.4	6.7	6.2	6.0	6.4

Note: ID = identification number of participants, AI = accuracy of information, GC = Grammatical correctness, SD = speed of delivery, CP = control of pauses, CD = control of other disfluencies, OF = overall fluency

Table 3. Pearson's *r* correlations between accuracy-related and fluency-related ratings.

	Speed of delivery	Control of pause	Control of disfluencies	Overall fluency
Accuracy of information	0.88**	0.86**	0.92**	0.92**
Accuracy of grammar	0.90**	0.85**	0.86**	0.88**

Note: ** p < .01 (two-tailed)

5.2 *Acoustic measures of fluency*

In this section, the fourteen temporal and melodic variables are presented (see Table 4). Table 4 also shows values of the different temporal and melodic variables for delivery 1 *vs* delivery 3. The D3/D1 ratio is the mean of each acoustic variable for delivery 3, divided by that for delivery 1.

For ten of the fourteen acoustic variables, scores changed in accordance with a more favorable fluency rating; the remaining four parameters were unaffected by the repeated delivery.~~For most of the acoustic variables, scores were higher in delivery 3~~. These exceptions are mean length of filled pauses, phonation/time ratio, median pitch and size of effective pitch range which hardly changed. The number of filled pauses and number of disfluencies were halved in delivery 3. Overall, the temporal measures of fluency were consistent with the trends for the fluency ratings in our earlier study, where both beginners and advanced students achieved significantly higher scores in delivery 3 (Yu & van Heuven 2013).

In Section 5.3, the relations between the quantitative prosodic measures and the fluency ratings will be explored in greater detail.

5.3 *Correlations between quantitative prosodic measures and fluency ratings*

In this section, the quantitative prosodic measures are compared with the fluency ratings so as to determine how, and to what extent, the two were related. Pearson's *r* ~~results~~ values are shown in Table 5. Twelve out of the fourteen quantitative prosodic measures were closely correlated with the judged fluency. Effective speech rate had the highest correlation (r = .84, p < .01), followed by mean length of fluent runs (r = .78, p < .01) and phonation/time ratio (r = .78, p < .01).

Table 4. Descriptive statistics of 14 acoustic fluency measures for 12 participants in two consecutive interpretations.

Variable		BA students							MA students					Mean	D3/D1
	Del.	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**		
Articulation rate (syll/s)	1	5.2	4.5	5.7	3.6	3.6	3.2	4.4	5.1	4.8	4.8	5.4	4.3	4.6	1.1
	3	5.1	3.8	4.7	4.4	4.5	4.0	5.0	5.1	5.0	5.0	5.0	6.2	4.8	
Speech rate (syll/s)	1	3.2	2.3	3.1	2.3	2.3	2.1	2.0	3.7	3.7	4.0	3.7	2.9	2.9	1.2
	3	3.8	2.3	3.3	3.1	3.5	2.8	3.2	4.1	3.8	3.9	3.7	3.8	3.4	
Effective speech rate (syll/s)	1	2.8	1.8	2.8	1.7	2.2	1.9	1.7	3.7	3.7	3.9	3.5	2.6	2.7	1.2
	3	3.3	2.2	3.3	2.9	3.4	2.8	2.9	4.1	3.7	3.8	3.6	3.4	3.3	
N silent pauses	1	42	95	54	61	60	43	63	36	34	28	54	6	53	.7
	3	41	59	35	32	38	30	45	27	33	29	42	54	39	
Length of silent pauses (s)	1	.7	.5	.7	.6	.7	1.0	1.1	.5	.4	.4	.4	.5	.6	.8
	3	.4	.6	.7	.5	.4	.8	.7	.4	.5	.5	.4	.5	.5	
N filled pauses	1	15	61	6	35	9	11	25	5	6	0	3	14	16	.5
	3	13	40	0	9	2	3	7	5	3	0	6	11	8	
Length of filled pauses (s)	1	.3	.5	.2	.4	.4	.3	.4	.3	.2	.0	.4	.3	.3	1
	3	.3	.4	.0	.5	.4	.3	.2	.3	.3	.0	.3	.3	.3	
N pauses	1	57	156	60	96	69	54	88	41	40	28	57	82	69	.7
	3	54	99	35	41	40	33	52	32	36	29	48	65	47	
Length of pauses (s)	1	.6	.5	.7	.5	.6	.8	.9	.5	.4	.4	.4	.5	.6	.8
	3	.4	.5	.7	.5	.4	.8	.6	.4	.5	.5	.4	.5	.5	
No. disfluencies	1	12	17	7	28	8	8	16	2	3	2	15	9	11	.5
	3	14	4	3	9	3	2	7	0	3	3	4	11	5	
Length of fluent runs (syllables)	1	4.8	2.2	4.7	3.0	4.1	4.9	3.2	6.9	6.6	9.5	5.3	4.2	5.0	1.3
	3	5.8	2.9	7.3	5.7	6.7	7.1	5.2	8.2	7.6	9.9	5.5	4.8	6.4	
Phonation/time ratio	1	.6	.5	.6	.6	.7	.7	.5	.7	.8	.8	.7	.7	.7	1
	3	.7	.6	.7	.7	.8	.7	.6	.8	.8	.8	.7	.6	.7	
Median pitch (Hz)	1	205	220	235	209	290	229	242	134	167	230	122	212	208	1
	3	203	220	247	220	296	236	252	136	169	223	119	208	211	
Effective pitch range (semitones)	1	3.8	4.6	7.5	5.2	7.6	4.5	4.7	5.1	8.4	6.5	9.4	6.3	6.1	1
	3	4.6	4.5	7.3	5.1	6.5	4.1	5.1	4.7	8.8	6.9	8.9	6.3	6.1	

Table 5. Pearson's correlation between fluency ratings and acoustic fluency measures for 12 participants in two consecutive interpretations

	Accuracy of information	**Grammatical correctness**	**Speed of delivery**	**Control of pause**	**Control of disfluencies**	**Overall fluency**
Articulation rate (syll/s)	0.27	0.34	0.50*	0.42*	0.33	0.36
Speech rate (syll/s)	0.65**	0.66**	0.83**	0.84**	0.74**	0.77**
Effective speech rate (syll/s)	0.72**	0.70**	0.86**	0.90**	0.82**	0.84**
N silent pauses	−0.50*	−0.54**	−0.58**	−0.68**	−0.64**	−0.60**
Length of silent pauses (s)	−0.56**	−0.42*	−0.62**	−0.64**	−0.51*	−0.62**
N filled pauses	−0.33	−0.37	−0.43*	−0.55**	−0.50*	−0.42*
Length of filled pauses (s)	−0.44*	−0.52**	−0.56**	−0.50*	−0.52**	−0.52**
N pauses	−0.44*	−0.49*	−0.53**	−0.65**	−0.60**	−0.54**
Length of pauses (s)	−0.54**	−0.43*	−0.62**	−0.59**	−0.46*	−0.58**
N disfluencies	−0.57**	−0.56**	−0.62**	−0.71**	−0.73**	−0.69**
Length of fluent runs (syll)	0.66**	0.68**	0.73**	0.82**	0.82**	0.78**
Phonation/time ratio	0.68**	0.63**	0.72**	0.83**	0.77**	0.78**
Median pitch (Hz)	−0.43*	−0.22	−0.45*	−0.37	−0.34	−0.41*
Effective pitch range (semitones)	0.39	0.22	0.37	0.42*	0.39	0.34

Note: * p < .05, ** p < .01 (two-tailed)

5.4 *Quantitative prosodic measures as predictors of judged fluency*

Several linear regression models were built in stepwise mode, using SPSS, to investigate to what extent the fourteen quantitative prosodic measures could explain the variance in fluency ratings. Tables 6, 7, 8 and 9 show the adjusted proportion of variance explained (R^2) by these models, and thus the incremental predictive power of each acoustic parameter.

First, model 1 (Table 6) evaluates all fourteen quantitative prosodic measures as predictors of fluency ratings: the adjusted R^2 shows that 78.9% of the variance in the ratings on speed of delivery may be explained on the basis of two temporal measures – i.e., effective speech rate (R^2 = 72.1%) and number of filled pauses (R^2 = 6.8%). Effective speech rate appeared to be the best indicator of the ratings on speed of delivery. Second, 88.2% of the variance in the ratings for control of pauses (Table 7) may be explained on the basis of three temporal measures – i.e., effective speech rate (R^2 = 79.6%), articulation rate (R^2 = 4.5%) and number of filled pauses (R^2 = 4.1%). Again, effective speech rate appeared to be the best indicator here. Third, model 3 (Table 8) shows that 87.6% of the variance in the ratings on control of disfluencies may be explained on the basis of four temporal measures – i.e., effective speech rate (R^2 = 66.1%), speech rate (R^2 = 9.1%), number of filled pauses (R^2 = 6.6%) and mean length of fluent run (R^2 = 5.8%). Here, too, effective speech rate appeared to be the best indicator. Finally, model 4 (Table 9) shows that 90.2% of the variance in the overall fluency ratings may be explained on the basis of four temporal measures – i.e., effective speech rate (R^2 = 68.7%), number of filled pauses, (R^2 = 6%), articulation rate (R^2 = 11.8%) and mean length of pause (R^2 = 3.7%). Once again, effective speech rate appeared to be the best predictor of judged overall fluency.

Table 6. Model 1 (dependent variable: judged speed of delivery)

Predictors	R^2	Adj R^2	increment	SE of estimate
effective speech rate	.734	.721		.53479
number of filled pauses	.807	.789	.068	.46557

Table 7. Model 2 (dependent variable: judged control of pauses)

Predictors	R^2	Adj R^2	increment	SE of estimate
effective speech rate	.805	.796		.45052
articulation rate	.855	.841	.045	.39763
number of filled pauses	.897	.882	.041	.34262

Table 8. Model 3 (dependent variable: judged control of disfluencies)

Predictors	R^2	Adj R^2	increment	SE of estimate
effective speech rate	.676	.661		.55945
speech rate	.773	.752	.091	.47875
number of filled pauses	.842	.818	.066	.40961
mean length of fluent runs	.897	.876	.058	.33850

Table 9. Model 4 (dependent variable: judged overall fluency)

Predictors	R^2	Adj R^2	increment	SE of estimate
effective speech rate	.701	.687		.56339
number of filled pauses	.769	.747	.060	.50647
articulation rate	.883	.865	.118	.36983
mean length of pauses	.919	.902	.037	.31471

5.5 A closer examination of the melodic parameters

As the two melodic parameters were not found to correlate as closely with the judged fluency ratings as the twelve temporal measures, we examined the relationship between judged fluency and each of the two melodic parameters separately. Figure 2 plots judged overall fluency as a function of the effective pitch range (in semitones, in panel A) and of the median pitch (in Hz, in panel B) for our 12 participants. In each panel the data points were plotted separately for the first (circles) and the third (triangles) delivery by the participant.

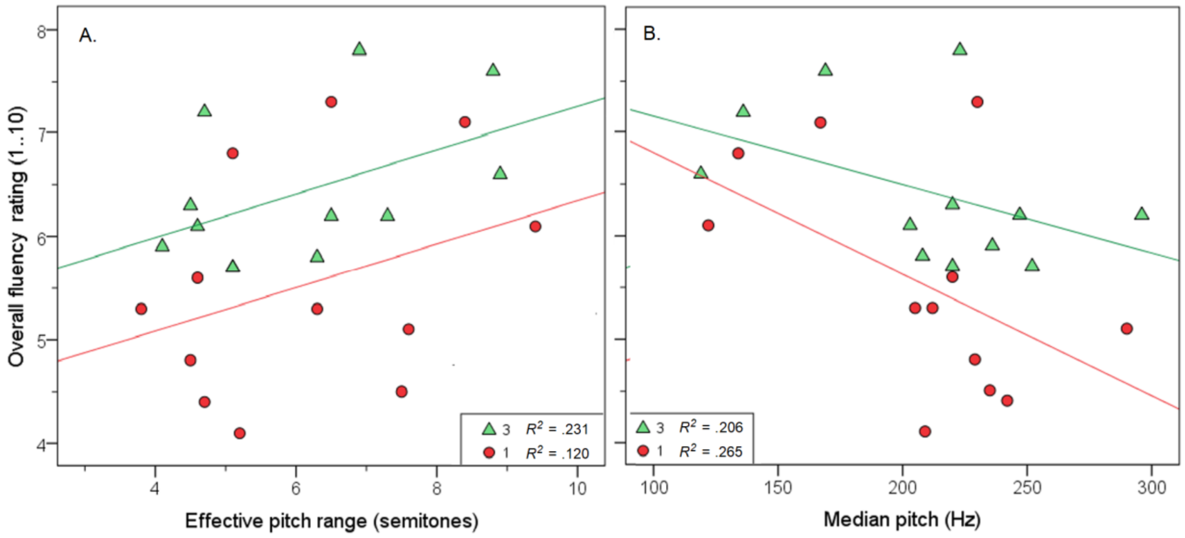

Figure 2. Judged overall fluency as a function of effective pitch range (semitones, panel A) and of median pitch (Hz, panel B) for the twelve participants broken down by delivery.

Figure 2A shows that there was a weak (almost negligible) correlation between the effective pitch range and judged fluency during the first delivery. A stronger correlation was seen in the third (and better) delivery, suggesting that the raters tended to associate larger pitch movements with speech confidence and competence, which perhaps may affect their perception of fluency in interpreting. However, the effective pitch range did not change systematically with delivery, *t* (11) = .449, *p* = .331 (one-tailed). But the interpreters with a larger effective pitch range were judged to be more fluent when the fluency judgments and the ranges were averaged across the two deliveries, *r* = .416, *p* = .089 (one-tailed).

Figure 2B shows that there was a moderate and negative correlation between median pitch and judged overall fluency, indicating that low-pitched interpreters were perceived as more confident and competent, which in turn, might also affect raters' perception of fluency in interpreting. Similarly, pitch level did not change systematically with delivery, *t* (11) = 1.552, *p* = .075 (one-tailed). The overall trend to judge the interpreters with lower medium pitch as being more fluent was significant, *r* = −.509, *p* = .049 (one-tailed).

The interpreters with the highest fluency ratings were typically the MA students. One of the five MA participants was a male speaker, who, naturally, has a lower median pitch than the eleven female speakers. Crucially, however, when we omitted the single male speaker from the analysis, the correlational strength between median pitch and judged fluency of interpreting was not affected, and in fact increased somewhat, $r = -.536$, $p = .044$ (one-tailed), which indicates that the association between low (median) pitch and judged fluency (mediated by the attribution of competence) was not confounded by the potential gender effect.

6. Discussion

Similar to what we found earlier (Yu & van Heuven 2017), all four judged fluency criteria (speed of delivery, control of pauses, control of disfluencies, overall fluency) correlated significantly with almost all the twelve temporal measures of fluency, despite that judged control of disfluencies and judged overall fluency did not correlate significantly with articulation rate (see Table 5). In particular, the study indicates that effective speech rate, phonation/time ratio, length of fluent runs, and speech rate had fairly strong correlations with perceived fluency ratings ($r \geqslant 0.77$, $p<0.01$), which generally corroborates the findings of other studies of similar kind (Yang 2015; Han et al. 2020). However, the overall interconnection between perceived fluency ratings and temporal measures including measures of speed fluency, breakdown fluency and repair fluency in this study was not found in previous research (e.g., Yang 2015; Han et al. 2020), in which only measures of speed fluency such as speech rate and measures of pausing such as mean length of silent pauses were found to be useful correlates of perceived fluency. In this case, further empirical studies will be needed to provide more ample evidence.

The two newly added melodic parameters were mainly found to correlate weakly or moderately with the four judged fluency criteria: median pitch correlated significantly (but moderately) with judged speed of delivery and judged overall fluency; and size of effective pitch range correlated significantly (but moderately) with judged control of pause (Table 5). The result resonates with findings of previous research (Collados Aís et al. 2007 and Christodoulides & Lenglet 2014), suggesting that interpreting-specific prosodic features are associated with the perception of fluency. It is also worth noting that, unlike the twelve temporal measures of fluency, the two melodic parameters did not change systematically with delivery, meaning that both the pitch level and effective pitch range may not improve with repetition in interpreting.

The study also identified which of the fourteen quantitative prosodic measures could best predict the fluency ratings. The results of the linear regression models in terms of useful predictors of judged fluency are largely consistent with those found in our previous study (Yu & van Heuven 2017). Effective speech rate (i.e., number of syllables, excluding disfluencies, divided by the total duration of speech production and pauses) appeared to be the best predictor of all four judged fluency criteria in CI (Tables 5-8) which might be explained by the fact that effective speech rate incorporates three aspects of fluency (speed fluency, breakdown fluency and repair fluency). The other temporal measures related (albeit less closely) to the fluency ratings were number of filled pauses, articulation rate, and mean length of pause. However, the two melodic parameters, i.e. median pitch and size of effective pitch range, were not among the predictors of judged fluency in this study, although previous studies suggest that confident/dominant speakers have lower (mean) pitch (Ohala 1983) and execute larger pitch movements (Gussenhoven 2004). This might be attributed to one of the following reasons: (1) Compared with temporal measures of fluency that are sensitive to the varying cognitive effort involved in complex speech tasks such as CI, melodic features are relatively stable as a result of long-term training and thus less prone to alteration, even when the cognitive load is reduced via repetition; (2) Effects of the melodic parameters on judged fluency were only measured with global parameters in this study. We therefore cannot rule out the possibility that specific local pitch differences may correlate with perceived competence and/or fluency of interpreting.

The exploration of quantitative parameters of judged fluency in interpreting offers insights into what features of an interpretation potentially contribute to judged fluency. The results of this experimental study offer implications for the automatic assessment of fluency in interpreting performance. The advent of artificial intelligence for scoring spoken language texts (Litman, Strik, & Lim, 2018) opens up new possibilities for fluency assessment in interpreting to be delivered with precision, consistency and objectivity. Given the availability of technologies that can easily detect syllables, pauses and disfluencies by running relevant scripts in speech analysis software package PRAAT and that can very quickly calculate these objective measures, it is likely that the labor-intensive and impressionistic rating of CI exams could be facilitated by quantitative measurement of effective speech rate or a combination of the good fluency predictors identified. With the empirical results of this study together with those of other studies (e.g., Han et al. 2020), automatic scoring systems such as SpeechRater (for scoring spontaneous non-native speech in the context of the TOEFL iBT Practice Online) and Pearson's Ordinate technology (for scoring the spoken portion of PTE Academic) may hopefully find their application in interpreting assessment. This could be quite efficient, at least in terms of screening out candidates who do not score satisfactorily on major objective fluency measures useful in predicting human ratings. Our study should thus serve as an initial step towards the development of an automatic quantitative assessment tool for fluency in interpreting.

The empirical results may also have implications for informing and benefiting the design of rubrics of interpreting fluency assessment. For raters who are gauging fluency, perhaps the rubrics should emphasize effective speech rate, number of filled pauses, articulation rate, and mean length of pauses over the other objective fluency measures that did not contribute to explaining variance in the ratings and were therefore considered as poor empirical correlates in our preliminary study. However, further research of larger scale that draws on the state-of-the-art knowledge of language testing and other disciplines will be needed to identify the most ideal scalar descriptors for the evaluation of interpreting fluency.

Furthermore, the findings of this study are expected to shed light on facilitating formative assessment in interpreting training. More specifically, interpreting teachers may need to guide trainees to develop a more comprehensive understanding of an integrated concept of fluency, incorporating not only speed fluency but also breakdown and repair fluency. Our experiment shows that the number of filled pauses and the number of disfluencies were halved in delivery 3 over delivery 1 (Table 4). Two aspects of fluency (i.e., breakdown fluency and repair fluency, or what Grosjean referred to as secondary fluency variables) thus seem to have a more important effect on fluency ratings in CI than was the case in previous research on L2 read and spontaneous speech (Cucchiarini et al. 2000, 2002; Kormos & Dénes 2004; Pinget et al. 2013). In addition, interpreting trainees need to be oriented and instructed in the melodic aspects of their delivery, which proves to be significantly though not strongly correlated with judged fluency. As such, the objective prosodic information (temporal and melodic) generated (semi-)automatically may serve as evidence about trainees' learning progress, which is to be brought back to them through feedback provided by trainers, peers or trainees themselves.

7. Conclusion

This experimental study probes into the possibilities of quantitative assessment of fluency in CI by studying potential correlations between judged fluency and automatically quantified temporal and melodic measures. It was found that effective speech rate, phonation/time ratio, length of fluent runs, and speech rate had fairly strong correlations with perceived fluency ratings. The statistics of the regression models show that effective speech rate was the best predictor of perceived fluency followed by number of filled pauses, articulation rate, and mean length of pause. Melodic measures such as median pitch and size of effective pitch range were not found to contribute to the variance in fluency ratings.

The study has several limitations. Firstly, there seemed to be the concern of multicollinearity in both the four judged fluency criteria as well as the twelve acoustic temporal measures. In other words, the four judged fluency criteria might not be independent and mutually exclusive from each other and the same might be applied to the twelve acoustic fluency measures – e.g., it is possible that speed of delivery will be associated with both control of pauses and control of disfluencies. The overall fluency rating might be connected with the three partial criteria of fluency, while mean length of fluent runs might be related to the number of filled pauses and number of disfluencies. Therefore, the results of the relative contribution of the objective measures to explain variance in the fluency ratings should be interpreted with caution. Future research may need to circumvent the problem of multicollinearity by choosing variables that are, in theory and in practice, not too highly interrelated. The second limitation lies in our overly conservative approach to detect and categorize silent/filled pauses and repetitions as disfluencies, since the interpreters might deliberately use these for clarity and emphasis. As such, by drawing on work from applied linguistics, psycholinguistics, discourse analysis and sociolinguistics, more sophisticated manner of conceptualizing disfluencies should be developed in future studies (De Jong 2018), which should make an effort to distinguish (non-planned) disruptive disfluencies from planned speech pauses – which have been found to enhance communication (Scharpff & Van Heuven 1988; Van Heuven & Scharpff 1991; Scharpff 1994). Finally, there was lack of ecological validity, since no listeners or audiences were present in this study. This, of course, does not reflect the situational dynamics of CI in real-life practice. More ecologically valid data is needed in future studies to verify the findings obtained in this study.

To sum up, the present experimental study attempts to investigate the relations between subjective ratings of fluency and quantifiable prosodic measures in CI. The most powerful predictor(s) of perceived fluency might ultimately lend themselves to development of automatic assessment for trainee interpreters' fluency in an examination setting, inform and improve the designing of rubrics of interpreting fluency assessment as well as facilitate formative assessment. The results of the study are preliminary though, and future studies are expected to be more interdisciplinary-based, statistically more rigorous, of larger-scale and of real-life settings.

References

AIIC (2002). *Regulation governing admissions and language classification*. Geneva: AIIC.

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

Brumfit, C. (1984). *Communicative methodology in language teaching: The roles of fluency and accuracy.* Cambridge: Cambridge University Press.

Bühler, H. (1986). Linguistic (semantic) and extra-linguistic (pragmatic) criteria for the evaluation of conference interpretation and interpreters. *Multilingua* 5 (4), 231–235.

Chambers, F. (1997). What do we mean by fluency? *System* 25, 535–544.

Cheung, A. (2013). Non-native accents and simultaneous interpreting quality perceptions. *Interpreting* 15 (1), 25–47.

Chiaro, D. & Nocella, G. (2004). Interpreters' perception of linguistic and non-linguistic factors affecting quality: A survey through the World Wide Web. *Meta* 49 (2), 278–293.

Christodoulides, G., Lenglet, C. (2014). Prosodic correlates of perceived quality and fluency in simultaneous interpreting. Proc. 7th International Conference on Speech Prosody 2014, 1002-1006, DOI: 10.21437/SpeechProsody.2014-189.

Collados Aís, A. (1998/2002). Quality assessment in simultaneous interpreting: The importance of nonverbal communication. In F. Pöchhacker & M. Shlesinger (Eds.), *The interpreting studies reader*. London/New York: Routledge, 327–336.

Collados Aís, A., Pradas Macías, M., Stévaux E. & García Becerra, O. (Eds.) (2007). *La evaluación de la calidad en interpretación simultánea: Parámetros de incidencia*. Granada: Comares.

Cucchiarini, C., Strik, H. & Boves, L. (2000). Quantitative assessment of second language learners' fluency by means of automatic speech recognition technology. *Journal of the Acoustical Society of America* 107, 989–999.

Cucchiarini, C., Strik, H. & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *Journal of the Acoustical Society of America* 111, 2862–2873.

Jong, N. H. de & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods* 41 (2), 385-390.

Fillmore, C. J. (1979). On fluency. In C. J. Fillmore, D. Kempler & W. S. Wang (Eds.), *Individual differences in language ability and language behaviour*. New York: Academic Press, 85–101.

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

Gile, D. (1995). *Basic concepts and models for interpreter and translator training*. Amsterdam: John Benjamins.

Ginther, A., Dimova, S. & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing* 27 (3), 379–399.

Goldman-Eisler, F. (1968). *Psycholinguistics: Experiments in spontaneous speech*. New York:

Academic Press.

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

Grosjean, F. & 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.

Gussenhoven, C. (2004). *The Phonology of Tone and Intonation*. New York: Cambridge University Press.

Han, C., Chen, S-J, Fu, R-B., & Fan, Q. (2020). Modeling the relationship between utterance fluency and raters' perceived fluency of consecutive interpreting. *Interpreting*. doi: 10.1075/intp.00040.han

Heuven, V. J. van (1994). Introducing prosodic phonetics. In C. Odé & V. J. van Heuven (Eds.), *Experimental studies of Indonesian prosody*. Semaian 9. Leiden: Vakgroep Talen en Culturen van Zuidoost-Azië en Oceanië, Leiden University, 1-26.

Heuven, V. J. van (2017). Prosody and sentence type in Dutch. *Nederlandse Taalkunde*, 22 (1), 3–29, 44–46.

Kormos, J. & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System* 32, 145–164.

Heuven, V. J. van & P. J. Scharpff (1991). Acceptability of several speech pausing strategies in low quality speech synthesis: interaction with intelligibility. *Proceedings of the 12th International Congress of Phonetic Sciences*, Aix-en-Provence, 458–461.

Kurz, I. (1993). Conference interpretation: Expectations of different user groups. *The Interpreters' Newsletter* 5, 3–16.

Kurz, I. (2001). Conference interpreting: Quality in the ears of the user. *Meta* 46 (2), 394–409.

Kurz, I. (2003). Quality from the user perspective. In A. Collados Aís, M. Fernández Sanchez & D. Gile (Eds.), *La evaluación de la calidad en interpretación: Investigación*. Granada: Comares, 3–22.

Kurz, I. (2008). The impact of non-native English on students' interpreting performance. In G. Hansen, A. Chesterman & H. Gerzymisch-Arbogast (Eds.), *Efforts and models in interpreting and translation research*. Amsterdam: John Benjamins, 179–792.

Leeson, R. (1975). *Fluency and language teaching*. London: Longman.

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

Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency*. Michigan: The University of Michigan Press, 25–42.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Litman, D., Strik, H., & Lim, G. S. (2018). Speech technologies and the assessment of second language speaking: Approaches, challenges, and opportunities. *Language Assessment Quarterly* 15 (3), 294-309.

Mead, P. (2000). Control of pauses by trainee interpreters in their A and B languages. *The Interpreters' Newsletter* 10, 89–102.

Mead, P. (2002). Exploring hesitation in consecutive interpreting: An empirical study. In G. Garzone & M. Viezzi (Eds), *Interpreting in the 21st century: Challenges and* *opportunities*. Amsterdam: John Benjamins, 75–84.

Mead, P. (2005). Methodological issues in the study of interpreters' fluency. *The Interpreters' Newsletter* 13, 39–63.

Möhle, D. (1984). A comparison of the second language speech production of different native speakers. In H. W. Dechert, D. Möhle & M. Raupach (Eds.), *Second language productions*. Tübingen: Gunter Narr, 26–49.

Moser, P. (1996). Expectations of users of conference interpretation. *Interpreting* 1 (2), 145–178.

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

Ohala, J.J. (1983). Cross-language use of pitch: an ethological view. *Phonetica* 40, 1-18.

Pinget, A., Bosker, H. R., Quené, H., Sanders, T. & De Jong, N. H. (2014). Native speakers' perceptions of fluency and accent in L2 speech. *Language Testing* 31 (3), 349–365.

Pradas Macías, E. M. (2003). Repercusión del intraparámetro pausas silenciosas en la fluidez: Influencia en las expectativas y en la evaluación de la calidad en interpretación simultánea. PhD dissertation, University of Granada.

Pradas Macías, E. M. (2007). La incidencia del parámetro fluidez. In A. Collados Aís, M. Pradas Macías, E. Stévaux & O. García Becerra (Eds.), *La evaluación de la calidad en interpretación simultánea: Parámetros de incidencia*. Granada: Comares, 53–70.

Pöchhacker, F. (1993). On the science of interpretation. *The Interpreters' Newsletter* 5, 52–59.

Pöchhacker, F. (2012). Interpreting quality: Global professional standards? In W. Ren (Ed.), *Interpreting in the age of globalization: Proceedings of the 8th National Conference and International Forum on Interpreting.* Beijing: Foreign Language Teaching and Research Press, 305–318.

Rennert, S. (2010). The impact of fluency on the subjective assessment of interpreting quality. *The Interpreters' Newsletter* 15, 101–115.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of non-native speaker conversations. *Discourse Processes* 14, 423–441.

Sawyer, D. B. (2004). *Fundamental Aspects of Interpreter Education: Curriculum and Assessment*. Amsterdam & Philadelphia: John Benjamins.

Scharpff, P. J. (1994). Het effect van spreekpauzes op de herkenning van woorden in voorgelezen zinnen [The effect of speech pauses on the recognition of words in read-out sentences], PhD dissertation, Leiden University.

Scharpff, P. J. & van Heuven, V. J. (1988). Effects of pause insertion on the intelligibility of low quality speech. In W.A. Ainsworth & J.N. Holmes (Eds.), *Proceedings of the 7th FASE/Speech-88 Symposium*. The Institute of Acoustics, Edinburgh, 261–268.

Schmidt, R. (1992). Psychological mechanisms underlying second language fluency. *Studies in Second Language Acquisition* 14, 357–385.

Shlesinger, M. (1994). Intonation in the production and perception of simultaneous interpretation. In S. Lambert & B. Moser-Mercer (Eds.), *Bridging the Gap: Empirical Research in Simultaneous Interpretation*. Amsterdam and Philadelphia, John Benjamins, 225–236.

Tavakoli, P. & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language*. Amsterdam: John Benjamins, 239–273.

Towell, R., Hawkins, R. & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied Linguistics* 1, 84–119.

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

Yang, L-Y. (2015). 中国口译学习者汉英交替传译 流利度的探索性研究 [An exploratory study of fluency in English output of Chinese consecutive interpreting learners]. *Journal of Zhejiang International Studies University* (1), 60–68.

Yu, W. T. & van Heuven, V. J. (2013). Effects of immediate repetition at different stages of consecutive interpreting: An experimental study. In: S. Aalberse & A. Auer (Eds.), *Linguistics in the Netherlands 2013.* Amsterdam: John Benjamins, 201–213.

Yu, W. T. & van Heuven, V. J. (2017). Predicting judged fluency of consecutive interpreting from acoustic measures Potential for automatic assessment and pedagogic implications. *Interpreting* 19 (1), 47–69.

Zhou, D. (2006). A study on the effects of input frequency and output frequency. *Modern Foreign Languages* 29, 154–163.

Zwischenberger, C. & Pöchhacker, F. (2010). Survey on quality and role: Conference interpreters' expectations and self-perceptions. *Communicate!* http://www.aiic.net/ViewPage.cfm/article 2510.htm (accessed 21 January 2013).

Funding

This research was supported by a Grant for the "*Prosodic Aspects in Chinese to English and English Chinese Consecutive Interpreting Training"* project sponsored by Shanghai International Studies University

Authors' address

Wenting Yu
Room 320, Building 1
550 Dalian Xi Road
Shanghai 200083
China

wenting_yu@163.com

About the authors

Wenting Yu teaches interpreting for the translation and interpretation majors at Shanghai International Studies University (SISU). She obtained her PhD degree in Linguistics from SISU in 2012. She was a visiting post-doctoral researcher at Leiden University Center for Linguistics (2012-2013). She specializes in interdisciplinary studies of interpreting, prosody, psycholinguistics and second language acquisition. Her publications include work on cognitive processing in interpreting, and interpreter assessment and training.

Vincent van Heuven is an emeritus professor of Experimental Linguistics and Phonetics at Leiden University, and at the University of Pannonia (Veszprém, Hungary). He served as Director of the Holland Institute of Linguistics and the Leiden University Centre for Linguistics, Chair of the Netherlands Graduate School in Linguistics, and Vice-President/Secretary of the Permanent Council of the International Phonetics Association. A former (associate) editor of several international book series and professional journals, he is a member of the Royal Netherlands Academy of Arts and Sciences.
V.J.J.P.van.Heuven@hum.leidenuniv.nl