What makes speech sound fluent? The contributions of pauses, speed and repairs

Hans Rutger Bosker, Anne-France Pinget, Hugo Quené, Ted Sanders and Nivja H de Jong

Language Testing 2013 30: 159 originally published online 6 October 2012
DOI: 10.1177/0265532212455394

Article

LANGUAGE TESTING

Language Testing
30(2) 159–175

DOI: 10.1177/0265532212455394
ltj.sagepub.com
SAGE

What makes speech sound fluent? The contributions of pauses, speed and repairs

Hans Rutger Bosker, Anne-France Pinget, Hugo Quené, Ted Sanders and Nivja H de Jong

Utrecht Institute of Linguistics OTS (UiL OTS), Utrecht University, The Netherlands

Abstract

The oral fluency level of an L2 speaker is often used as a measure in assessing language proficiency. The present study reports on four experiments investigating the contributions of three fluency aspects (pauses, speed and repairs) to perceived fluency. In Experiment 1 untrained raters evaluated the oral fluency of L2 Dutch speakers. Using specific acoustic measures of pause, speed and repair phenomena, linear regression analyses revealed that pause and speed measures best predicted the subjective fluency ratings, and that repair measures contributed only very little. A second research question sought to account for these results by investigating perceptual sensitivity to acoustic pause, speed and repair phenomena, possibly accounting for the results from Experiment 1. In Experiments 2–4 three new groups of untrained raters rated the same L2 speech materials from Experiment 1 on the use of pauses, speed and repairs. A comparison of the results from perceptual sensitivity (Experiments 2–4) with fluency perception (Experiment 1) showed that perceptual sensitivity alone could not account for the contributions of the three aspects to perceived fluency. We conclude that listeners weigh the importance of the perceived aspects of fluency to come to an overall judgment.

Keywords

Fluency perception, pauses, perceptual sensitivity, repair, speed

The level of oral fluency of non-native (L2) speakers is an important measure in assessing a person's language proficiency. It is often examined using professional tests (e.g. TOEFL iBT) which may have lasting effects on a person's life in the non-native cultural environment (such as employment or university admission). Therefore, researchers have

Corresponding author:
Hans Rutger Bosker, Utrecht University, Trans 10, Kamer 1.61, 3512 JK UTRECHT, The Netherlands.
Email: h.r.bosker@uu.nl

attempted to unravel the different factors that influence fluency ratings. Two different interpretations of the notion 'fluency' have been distinguished by Lennon (1990): fluency in the *broad* and in the *narrow* sense. Fluency in a *broad* sense is most often used in everyday life when, for instance, someone claims to be 'fluent' in French. In this setting, speaking a language fluently may refer to error-free grammar, a large vocabulary and/or native-like pronunciation. Fluency in the broad sense is equivalent to overall speaking proficiency (Chambers, 1997) and has been further categorized in Fillmore (1979). In contrast, fluency in a *narrow* sense is a component of speaking proficiency. This sense is often encountered in oral examinations: apart from grammar and vocabulary, the flow and smoothness of the speech is also assessed. Fluency in this sense has been defined as an 'impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and smoothly' (Lennon, 1990, p. 391) and it is this narrow sense that we are concerned with here.

Segalowitz (2010) has, more recently, approached this fluency in the narrow sense from a cognitive perspective. He argues that sociolinguistic (social context), psycholinguistic (the neurocognitive system of speech production) and psychological (motivation) factors interlinked in a dynamical system all contribute to the level of fluency. Three facets of fluency are distinguished, namely *cognitive fluency* – 'the efficiency of operation of the underlying processes responsible for the production of utterances'; *utterance fluency* – 'the features of utterances that reflect the speaker's cognitive fluency' which can be acoustically measured; and *perceived fluency* – 'the inferences listeners make about speakers' cognitive fluency based on their perceptions of their utterance fluency' (Segalowitz, 2010, p.165). Furthermore, measures of utterance fluency (e.g. number and duration of filled and silent pauses, speech rate, number of repetitions and corrections, etc.) may be clustered into three fluency aspects: *breakdown fluency* concerns the extent to which a continuous speech signal is interrupted; *speed fluency* has been characterized as the rate and density of speech delivery; and *repair fluency* relates to the number of corrections and repetitions present in speech (Skehan, 2003, 2009; Tavakoli & Skehan, 2005).

The present study investigates the separate contributions of these latter fluency aspects to perceived L2 fluency. This issue is approached from two perspectives: from the language testing perspective (Experiment 1) and from a cognitive psychological perspective (Experiments 2, 3, 4). Many previous studies have looked at factors influencing raters' judgments (e.g. Iwashita et al., 2008); the present study is an attempt to extend this body of research by relating subjective fluency ratings of L2 speech to combinations of acoustic measures, specific to each of the three fluency aspects. In this fashion we intend to determine the relative contributions of the fluency aspects to perceived L2 fluency (Experiment 1). Once this has been established, the question of *why* some fluency aspects contribute more to fluency perception than others will be addressed. To answer this question, we turn to cognitive psychological factors. More specifically, we hypothesize that listeners' general perceptual sensitivity lies at the foundation of fluency perception. A series of experiments aims to establish the relative sensitivity of listeners to pause phenomena (Experiment 2), to the speed of delivery (Experiment 3) and to repair features in speech (Experiment 4). Results of such

investigations license a comparison between listeners' sensitivity to speech characteristics and the factors involved in L2 fluency perception. This comparison is expected to shed light on the question of *why* some fluency aspects contribute more to fluency perception than others.

The approach of our experiments involves relating utterance fluency (objective phonetic measurements of L2 speech) to perceived fluency (subjective ratings of the same speech). This approach is often used to gain more insight into the acoustic correlates of oral fluency. For instance, Cucchiarini, Strik & Boves (2002) had teachers rate speech materials obtained from 30 beginning learners and 30 intermediate learners of Dutch. These perceived fluency ratings were found in subsequent analyses to be best predicted by the *number of phonemes per second* for beginning learners and by the *mean length of run* for the intermediate learners. Derwing et al. (2004) used novice raters for obtaining perceived fluency judgments. These raters listened to speech materials of 20 beginner Mandarin-speaking learners of English. Significant correlations were found between the fluency ratings and *pausing* and *standardized pruned syllables per second* (the total number of syllables disregarding corrections, repetitions, non-lexical filled pauses, etc.). Rossiter (2009) found *number of pauses per second* and *pruned speech rate* to be strong predictors of perceived fluency. Kormos and Dénes (2004) related acoustic measurements from L2 Hungarian speakers to fluency ratings by native and non-native teachers. They found *speech rate, mean length of utterance, phonation time ratio* (spoken time / total time × 100%) and *the number of stressed words produced per minute* to be the best predictors of fluency scores.

A closer look into the methodology and results of these studies reveals much diversity. Conceptual considerations have major effects on the studies' designs and results. To illustrate this point, consider the intercollinearity of acoustic measures of speech. Depending on the specificity of speech annotations, the number of available acoustic predictors of speaking fluency may grow very large. The larger the number of acoustic measures that are related to fluency ratings, the larger the chance of confounding the different measures, which would obscure the interpretability of results. For example, the measures of *speech rate* (number of syllables divided by total time including silences) and *mean duration of a silent pause* both depend on the duration of silent pauses in the speech signal, and therefore, these two measures are interrelated. If a study found these two measures to be strongly related to fluency ratings, the relative contribution of each measure to perceived fluency would remain unclear, owing to the intercollinearity of these measures. In order to understand what raters really listen to when evaluating oral fluency, correlations among acoustic measures should also be taken into account. Unfortunately, correlations between fluency measures are often lacking in the literature, even though the degree of intercollinearity of measures may distinguish orthogonal from confounded measures. De Jong et al. (in press) and Pinget et al. (forthcoming) did report correlations between acoustic measures, and argued that using measures with low intercollinearity aids the interpretability of results. The present study also emphasizes the degree of intercollinearity of our measures. More specifically, the distinction between the three fluency aspects (*breakdown, speed* and *repair fluency*) is central to our selection of acoustic

measures. Only those measures that do not confound the fluency aspects will be employed in our regression analyses.

The first experiment of this study was set up to answer a first research question:

> Research question 1: Which of the fluency aspects (breakdown, speed or repair fluency), *as represented by sets of acoustic measures*, is most related to perceived fluency?

This issue is approached by relating objective acoustic measurements of speech to subjective fluency ratings of that same speech. A group of untrained raters judged the fluency of L2 Dutch speech excerpts. Derwing et al. (2004) already hypothesized that fluency judgments from untrained native-speaker raters are equivalent to those obtained from expert raters, given comparable levels of inter-judge agreement. Rossiter (2009) compared fluency ratings from untrained raters with fluency ratings from expert raters and did not find a statistically significant difference between the two groups. Also, Pinget et al. (forthcoming) have recently demonstrated that untrained raters can keep the concept of fluency well apart from perceived accent. The subjective ratings from the untrained raters from Experiment 1 were modeled by three sets of predictors: a set of pause measures, a speed measure and a set of repair measures. Since the discussed literature (e.g. Cucchiarini, Strik & Boves, 2002; Derwing et al., 2004; Kormos & Dénes, 2004; Rossiter, 2009) mainly found speed and pause measures to be related to fluency ratings, it is expected that both *breakdown* and *speed fluency* are primary factors influencing fluency ratings. With respect to *repair fluency*, the literature seems to suggest that there is no relationship between repair fluency and perceived fluency. For instance, Cucchiarini, Strik & Boves (2002) did not find any relationship between fluency ratings and *number of disfluencies* (which covers, among others, repetitions and corrections).

Experiment 1 is expected to shed light on RQ 1 by distinguishing the relative contributions of the three fluency aspects. Finding an answer to RQ 1 raises a second question of *why* some fluency aspects contribute more to fluency perception than others. To this end, the psycholinguistic process of speech perception is investigated. One specific cognitive psychological factor possibly underlying fluency perception is targeted, namely listeners' general perceptual sensitivity. Thus the relationship between the sensitivity of listeners to speech characteristics and fluency perception is studied. It is hypothesized that differences in sensitivity to specific speech phenomena may account for differences in correlations between acoustic measures and fluency ratings. More specifically, if, for instance, pause measures can be found to be strongly related to perceived fluency ratings, the question can be posed about whether this might be owing to the fact that listeners are in general more sensitive to pause phenomena. If this scenario can be shown to be true, perception then 'paves the way' for rating: the way we perceive speech influences our subjective impression of that speech. If, in contrast, there is an asymmetry between speech features that contribute to fluency perception and the features in speech that listeners are most sensitive to (e.g. pause characteristics are well perceived but contribute only a small amount to fluency perception), then perceptual sensitivity is not the only factor

determining fluency perception. Listeners, in this scenario, would first perceive the acoustic characteristics of a speaker's speech but then subsequently also weigh their importance for fluency. These considerations result in the formulation of our second research question:

> Research question 2: Which acoustic speech properties are listeners most sensitive to?

To answer RQ 2, three additional experiments were designed. The crucial distinction between the experiments was the set of instructions given to raters. In Experiment 2 the same L2 speech materials from Experiment 1 were used but a new group of raters received different instructions, namely to rate the use of silent and filled pauses. Relating their pause ratings to objective pause measures is expected to reveal to what extent listeners are sensitive to pauses in speech. Experiment 3 had a similar approach, but now another group of raters was instructed to rate the identical L2 speech materials on the speed of delivery of the speech. And in Experiment 4, yet another group of raters received instructions to rate the L2 speech on the use of repairs (i.e. corrections and hesitations). Findings from these latter three experiments allow us to explore whether the different sensitivities of listeners to acoustic speech characteristics (RQ 2) may account for the relative contributions of fluency aspects to perceived fluency (RQ 1).

As mentioned above, the first research question is approached in Experiment 1 by relating *objective* acoustic measurements from three aspects of fluency to *subjective* ratings. Additional support for findings from Experiment 1 may be found by relating the *subjective* perception of the three fluency aspects (Experiments 2–4) to *subjective* ratings on fluency (Experiment 1):

> Research question 3: Does predicting fluency ratings with the *subjective* perception of fluency aspects (breakdown, speed or repair fluency) as predictors lead to similar results as when predicting fluency ratings using *objective* measures of the fluency aspects?

Instead of using objective acoustic details on pausing, speed and repairs in speech, we now use the subjective ratings (from Experiments 2–4) on these same dimensions as predictors for the perceived fluency ratings (from Experiment 1). If a similar hierarchy of fluency aspects can be established, then RQ 3 would yield extra support for findings from Experiment 1.

Method

Participants

Eighty participants, recruited from the UiL OTS subject pool, were paid for participation in one of four experiments. All were native Dutch speakers without any training in language rating and reported normal hearing (Experiment 1: $n = 20$, mean age = 20.20, $SD = 1.88$, 1m/19f; Experiment 2: $n = 20$, mean age = 20.65, $SD = 2.70$, 2m/18f; Experiment

3: n = 20, mean age = 20.35, SD = 2.76, 2m/18f; Experiment 4: n = 20, mean age = 20.74, SD = 1.79, 4m/16f).

Stimulus description

Speech recordings from native and non-native speakers of Dutch were obtained from the 'What Is Speaking Proficiency'-project (WISP) in Amsterdam (as described in De Jong et al., 2012). Assessment of these speakers' productive vocabulary knowledge resulted in vocabulary scores which were shown to be highly representative of their overall speaking proficiency (De Jong et al., 2012). Two non-native speaker groups (15 English and 15 Turkish) were matched on their performance on the vocabulary test (Turkish: mean score = 68, SD = 18; English: mean score = 64, SD = 16; $t(28) = 0.552$, $p = .585$). Moreover, eight native speakers of Dutch were also selected from the WISP corpus. These were included in order to offer raters reference points to which they could compare the non-native items. The native speakers were selected such that their vocabulary scores were closest to the average of all native speakers (= 106).

All speakers had performed eight different computer-administered speaking tasks. These tasks had been designed to cover the following three dimensions in a 2 × 2 × 2 fashion: *complexity* (simple, complex), *formality* (informal, formal) and *discourse type* (descriptive, argumentative). From these eight tasks three tasks were here selected that covered a range of task characteristics and targeted relatively long stretches of speech. In Table 1 descriptions of each task are given together with the proficiency level according to the Common European Framework of Reference for Languages (CEFR) (Hulstijn et al., 2012).

In this fashion, the speech materials consisted of 38 speakers performing three tasks (= 114 items). Fragments of approximately 20 seconds were excerpted from approximately the middle of the original recordings. Each fragment started at a phrase boundary (Analysis of Speech Unit; Foster, Tonkyn & Wigglesworth, 2000) and ended at a pause

Table 1. Descriptions of the selected speaker tasks.

	CEFR-level	Characteristics	Description
Task 1	B1	Simple, formal, descriptive	The participant, who has witnessed a road accident some time ago, is in a courtroom, describing to the judge what had happened.
Task 2	B1	Simple, formal, argumentative	The participant is present at a neighbourhood meeting in which an official has just proposed to build a school playground, separated by a road from the school building. Participant gets up to speak, takes the floor, and argues against the planned location of the playground.
Task 3	B2	Complex, formal, argumentative	The participant, who is the manager of a supermarket, addresses a neighbourhood meeting and argues which one of three alternative plans for building a car park is to be preferred.

(>250ms). The fragments had a sampling frequency of 44100Hz and were scaled to an intensity of 70dB.

Six objective acoustic measures were calculated for each recording (Table 2) based on human annotations of the speech recordings. Confounding the fluency aspects was avoided so that each measure was specific to one aspect of fluency. For this reason, all frequency measures were calculated using spoken time (excluding silences) instead of total time (including silences). For instance, previous work suggests that the measure *mean length of run* correlates with raters' perceptions of fluency (Cucchiarini, Strik & Boves, 2002; Kormos & Dénes, 2004), but because this measure is dependent on the number of pauses in speech it actually combines both speed and breakdown fluency. Therefore, this type of measure was not used in the present study. The aspect of speed fluency was represented by one measure: the mean length of syllables (MLS). A log transformation was performed so that the data would more closely approximate the normal distribution. Breakdown fluency was represented by three measures: the number of silent pauses per second spoken time (NSP), the number of filled pauses per second spoken time (NFP) and the mean length of silent pauses (MLP). A log transformation was performed also on this latter measure for the same reasons as above. These three measures were selected, since we wanted to have separate measures for the *number* and the *duration* of silent pauses, and since we wanted to make the distinction between filled and silent pauses. Finally repair fluency was represented by two measures: the number of repetitions (NR) and the number of corrections (NC) per second spoken time. All measures have the same polarity: the higher a value, the less fluent the fragment. The pause exclusion criterion was set at 250ms (Towell, Hawkins &Bazergui, 1996) since pauses shorter than 250ms can be classified as micro-pauses (Riggenbach, 1991) which are not regarded as hesitation phenomena.

Design and procedure of Experiment 1

The speech fragments of approximately 20 seconds long were presented to participants using the FEP experiment software (version 2.4.19, Veenker, 2006). Participants listened to stimuli over headphones at a comfortable volume in sound-attenuating booths. Written instructions, presented on the screen, instructed participants to judge the speech

Table 2. List of six selected acoustic measures.

Aspect	No.	Acoustic measures	Calculation
Speed	1	Mean length of syllables (MLS)	Log (spoken time / number of syllables)
Breakdown	2	Number of silent pauses (NSP)	Number of silent pauses / spoken time
	3	Number of filled pauses (NFP)	Number of filled pauses / spoken time
	4	Mean length of silent pauses (MLP)	Log (sum of length of silent pauses / number of silent pauses)
Repair	5	Number of repetitions (NR)	Number of repetitions / spoken time
	6	Number of corrections (NC)	Number of corrections / spoken time

Note: Spoken time = duration of speech fragment excluding silences of >250ms.

fragments on overall fluency. In order to avoid the interpretation of fluency in the broad sense (i.e. overall speaking proficiency), participants were instructed *not* to rate the items in this broad interpretation. In contrast, participants were asked to base their judgments on (1) the use of silent and filled pauses, (2) the speed of delivery of the speech and (3) the use of hesitations and/or corrections (and not on grammar, for example). Following the instructions but prior to the actual rating experiment six practice items were presented so that participants could familiarize themselves with the procedure. When participants posed questions to the experimenters, no other instructions than the written instructions were supplied to the participants by the experimenters. There were three different pseudo-randomized ordered lists of the stimuli and three reversed versions of these lists, resulting in six different orders of items. Each session lasted approximately 45 minutes. Participants were allowed to take a brief pause halfway through the experiment. Participants rated the speech fragments using an Equal Appearing Interval Scale (EAIS; Thurstone, 1928). This scale was composed of nine stars with labeled extremes ('not fluent at all' on the left; 'very fluent' on the right; see Appendix). Above each rating scale a question summarized the rating instructions. At the end of each session the participant filled out a short questionnaire which enquired about attitudes towards and exposure to L2 speech, the factors which the participants themselves thought had influenced them in their rating task (e.g. pauses, speed, repairs, grammar, vocabulary, etc.), and personal details.

Design and procedure of Experiment 2

The speech materials used in the second experiment were identical to those in Experiment 1. A new group of 20 raters participated in this second experiment. The procedure of this experiment was identical to Experiment 1, but crucially the instructions given to these new raters were altered. Participants in Experiment 2 were asked to rate the speech for the use of silent and filled pauses. The instructions to participants in Experiment 2 were modeled on those used for Experiment 1 (i.e. the introduction, specific formulations and the definitions of pause phenomena) but no reference was made to the notion of 'fluency'.

Design and procedure of Experiment 3

The speech materials and procedure of the previous experiments were used again for the third experiment. A new group of raters was instructed to rate the L2 speech with the instructions to base their judgments on the speed of delivery of the speech. The literal instructions were modeled on Experiment 1 such that certain terms and the definition of 'speed of delivery' were identical across experiments but without mentioning the term 'fluency'.

Design and procedure of Experiment 4

In the fourth experiment another group of raters was instructed to rate the same L2 speech materials on the use of hesitations and corrections. Again, definitions of repair phenomena were identical to Experiment 1 but no reference was made to the notion of 'fluency'.

Results

Acoustic analysis of stimulus materials

First of all, the non-native speech materials were analysed (no analysis was performed on (ratings on) native fragments). The intercollinearity of the acoustic measures was investigated through Pearson's *r* correlations between acoustic measures, in Table 3. The correlation measures reported in Table 3 allow a comparison between acoustic measures within and across aspects of fluency. It was only possible to analyse correlations within fluency aspects for breakdown and repair fluency, since speed fluency was represented by one single measure. Within breakdown fluency only one statistically significant correlation was found, namely a weak correlation between NSP and NFP ($r = -0.248$). Within repair fluency, the correlation between the two measures was not statistically significant. Correlations across fluency aspects primarily concerned weak to moderate correlations with the speed fluency measure MLS, but a correlation between NSP and NC was also found. The relationship between acoustic measures within fluency aspects was similar to the relationship between acoustic measures across fluency aspects.

In addition, correlations between single acoustic measures and the fluency ratings were calculated (see Table 3). The highest observed correlation was between the speed measure *mean length of syllables* and the fluency ratings ($r = -0.742$). In order to investigate the contribution of fluency aspects to perceived fluency, additional analyses were performed.

Results of Experiment 1

Each item in Experiment 1 was rated by 20 judges. The extent to which raters in Experiment 1 agreed with each other was high (Cronbach's alpha coefficient: 0.97). In

Table 3. Correlations (Pearson's *r*) between acoustic measures and between acoustic measures and fluency ratings.

Aspects	Acoustic measure	Speed	Breakdown			Repair		Fluency ratings
		MLS	NSP	NFP	MLP	NR	NC	
Speed	Mean length of syllables (MLS)	1						−0.742 ***
Breakdown	Number of silent pauses (NSP)	0.330 **	1					−0.422 ***
	Number of filled pauses (NFP)	0.308 **	−0.248 *	1				−0.154
	Mean length of silent pauses (MLP)	0.152	−0.096	−0.168	1			−0.470 ***
Repair	Number of repetitions (NR)	0.292 **	0.037	0.188	0.034	1		−0.348 ***
	Number of corrections (NC)	0.102	0.216 *	−0.037	−0.088	0.012	1	−0.241 *

Note: * = $p < .05$; ** = $p < .01$, *** = $p < .001$.

order to relate these subjective ratings on each item to the objective acoustic properties of that item, a method of collapsing these 20 ratings for each item was required. Many previous fluency studies take the mean of the 20 ratings for each item, thereby disregarding such confounding factors as individual differences between raters, for instance, or effects of presentation order. Our analyses were performed in two consecutive steps. The first step involved correcting the fluency ratings for these confounding factors using Best Linear Unbiased Predictors (Baayen, 2008, p. 247), which resulted in corrected *estimates* of the raw fluency ratings. The correction procedure was performed using Linear Mixed Models (cf. Quené & Van den Bergh, 2004, 2008; Baayen, Davidson & Bates, 2008) as implemented in the lme4 library (Bates & Maechler, 2010) in R (R Development Core Team, 2010). Thus we controlled for three confounding factors: Order (fixed effect) testing for general learning or fatigue effects; Rater (random effect) testing for individual differences between raters; and Order Within Raters (random effect) testing for individual differences in order effects. Simple models, containing one or two of these predictors, were compared to more complex models that contained one additional predictor. In order to allow such comparisons of models in our analysis, coefficients of models were estimated using the full Maximum Likelihood criterion (Pinheiro & Bates, 2000; Hox, 2010). Likelihood ratio tests (Pinheiro & Bates, 2000) showed that the most complex model proved to fit the data of Experiment 1 better than any simpler model. This optimal model showed significant effects of Rater, of Order (raters became harsher to the L2 speech as the experiment progressed) and of Order Within Raters (the order effect differed among individual raters). This optimal model was used to predict *estimates* of the fluency ratings. This was the first step of the investigative procedure reported here. All subsequent analyses were performed on these corrected estimates instead of on averages.

The second step involved relating objective acoustic measures to these corrected estimates of the fluency ratings. Multiple linear regression analyses were performed in order to explore to what extent a set of objective acoustic measures could explain the variance of the (estimated) fluency ratings, gauged by the adjusted R^2.

Because the present study is primarily concerned with the contributions of fluency aspects, and not of single acoustic measures, predictors in the multiple linear regression models were sets of acoustic measures and not single acoustic measures. All measures were centralized to their median value. In Table 4 six different models of the fluency judgments are summarized. Because effects of the L1 language (English vs. Turkish) and of the different speaking tasks were not statistically significant, these factors will be ignored in the present multiple linear regression analyses.

First of all, three models (1–3) were built with predictors from only one of the fluency aspects. Model (1) included the three acoustic measures specific to breakdown fluency: NFP, NSP and MLP. A comparison between a model with no interactions and a model with three two-way interactions demonstrated that the model with the three two-way interactions had a significantly stronger explanatory power and therefore these three two-way interactions were included in all subsequent models. This model resulted in an adjusted R^2 of 0.5917. Model (2) predicted fluency ratings using the speed measure MLS as predictor, and it resulted in an adjusted R^2 of 0.5449. Model (3) had the repair fluency measures, NC and NR, as predictors of perceived fluency (adjusted $R^2 = 0.1583$).

Table 4. Models predicting the fluency estimates of Experiment 1 using acoustic measures.

Model	Predictors	Adjusted R^2	Significance testing
(1)	NFP * NSP * MLP (breakdown)	0.5917	
(2)	MLS (speed)	0.5449	
(3)	NC + NR (repair)	0.1583	
(4)	NFP * NSP * MLP (breakdown) + MLS (speed)	0.7825	Model 4 vs. 1: $F(1,82) = 73.793$, $p < .001$
(5)	NFP * NSP * MLP (breakdown) + NC + NR (repair)	0.6804	Model 5 vs. 1: $F(2,81) = 12.523$, $p < .001$
(6)	NFP * NSP * MLP (breakdown) + MLS (speed) + NC + NR (repair)	0.8378	Model 6 vs. 4: $F(2,80) = 15.004$, $p < .001$

Seeing that model (1) with breakdown fluency measures as predictors explained the largest part of the variance of the fluency ratings, we tested whether additional contributions of speed fluency and of repair fluency added to the predictive power of the model. Model (4) additionally contained the acoustic measure specific to speed fluency, MLS (adjusted $R^2 = 0.7825$), and model (5) also included the repair fluency measures, NC and NR (adjusted $R^2 = 0.6804$). As evidenced by the higher adjusted R^2 values relative to model (1) and by the statistical comparisons of models, both models improved the explanatory power of model (1) with model (4) yielding a higher adjusted R^2 than model (5). Finally, the most complex model (6) which included all fluency aspects as predictors yielded the highest adjusted R^2 of 0.8378.

When comparing these results with the responses from the participants to the questions in the post-experimental questionnaire, it was found that participants themselves reported to have been mainly influenced by pauses ($n = 19$) and speed ($n = 15$) and less so by repairs ($n = 12$).

Results of Experiments 2–4

In Experiments 2–4 all stimulus material was kept constant, but new groups of raters received different instructions, namely to rate the speech on the use of silent and filled pauses (Experiment 2), on the speed of delivery (Experiment 3) and on the use of repetitions and corrections (Experiment 4). Raters within the separate experiments strongly agreed as evidenced by high Cronbach's alpha coefficients calculated on the raw ratings: 0.95 (Experiment 2); 0.96 (Experiment 3); 0.94 (Experiment 4). The analyses of the different experiments again involved two steps. Firstly, the raw ratings were corrected for confounding random effects. It was established that for all experiments the most complex Linear Mixed Model, which included Order, Rater and Order Within Raters as predictors, proved to fit the raters' data the best. The estimates resulting from these models were taken as dependent variable in the second step of the analyses. This second step involved modeling the subjective estimates of each experiment by objective measures from the appropriate fluency aspect (i.e. speed ratings by speed measures, pause ratings by pause measures, and repair ratings by repair measures). As given in Table 5, model

Table 5. Models predicting the estimates of Experiments 2–4 using acoustic measures.

Model	Dependent variable	Predictors	Adjusted R^2
(7)	Pause ratings from Experiment 2	NFP * NSP * MLP	0.6986
(8)	Speed ratings from Experiment 3	MLS	0.5287
(9)	Repair ratings from Experiment 4	NC + NR	0.5452

(7), predicting subjective pause ratings using pause measures, was observed to have the highest adjusted R^2 value (0.6986) of the three analyses. Models (8) and (9) perform worse than model (7) and explain almost the same amount of variance. The responses from the participants to the questions in the post-experimental questionnaire did not reveal any particular pattern, except that each group said to have been mainly influenced by the 'relevant' acoustic factor (e.g. pause raters by pauses, speed raters by speed, repair raters by repairs).

Subjective ratings as predictors for fluency ratings

The data resulting from Experiments 2–4 allow for an additional analysis of the results of Experiment 1 addressing RQ 3. Using the same materials, the subjective fluency ratings from Experiment 1 were predicted by the subjective ratings on specific speech characteristics from Experiments 2–4; see Table 6. These results show that most of the variance of the fluency judgments may be predicted by subjective pause ratings. The model with the 'best fit' was the most complex model (15), with the ratings on all three subjective dimensions included as predictors.

Discussion

This study investigated the contributions of three aspects of fluency (*breakdown, speed and repair fluency*) to perceived fluency ratings. In Experiment 1 untrained raters

Table 6. Models predicting the fluency judgments of Experiment 1 using subjective ratings.

Model	Predictors	Adjusted R^2	Significance testing
(10)	Pause estimates	0.8523	
(11)	Speed estimates	0.7829	
(12)	Repair estimates	0.2735	
(13)	Pause estimates + Speed estimates	0.8923	Model 13 vs. 10: $F(1,87) = 34.626$, $p < .001$
(14)	Pause estimates+ Repair estimates	0.8807	Model 14 vs. 10: $F(1,87) = 21.873$, $p < .001$
(15)	Pause estimates+ Speed estimates+ Repair estimates	0.9208	Model 14 vs. 13: $F(1,86) = 31.4$, $p < .001$

evaluated L2 speech items with regards to fluency, with the aim of establishing the contributions of the different fluency aspects to fluency perception (RQ 1). Sets of acoustic measures relating one of three fluency aspects were included in models predicting the subjective fluency ratings. Cross-correlations between the speech measures demonstrated that both within and across fluency aspects our speech measures were largely independent. This low intercollinearity aided the interpretation of other analyses. De Jong et al. (in press) also report on correlations between acoustic measures within and across fluency aspects. A comparison reveals that the relationship between measures that theoretically cluster together within fluency aspects show, in both studies, no stronger correlations amongst each other than measures across fluency aspects do. Together with De Jong et al. (in press) we argue that measures from the same fluency aspect might be caused by the same cognitive problems in the speech production process. Where one speaker would use a silent pause to win time, another might resort to the use of filled pauses, resulting in low correlations between the two measures. Future research into the specific function of disfluencies in (L1 and L2) natural speech will have to address this issue.

Having established that the acoustic measures used in our analyses did not confound the fluency aspects, we turn to RQ 1. Comparisons between fluency models reveals that all three aspects play a role in fluency perception and none of these aspects should be disregarded. Still, breakdown fluency explained the largest part of the variance in subjective fluency ratings, closely followed by speed fluency. Strong correlations between pause and speed measures and fluency ratings as reported in previous literature (e.g. Derwing et al., 2004; Rossiter, 2009) support this major role of breakdown and speed fluency. In addition, correlations between single acoustic measures and the fluency ratings suggest that the major role of breakdown fluency is primarily owing to the effect of (the duration and the number of) silent pauses rather than filled pauses.

The second research question sought to find a possible explanation for this finding by investigating the cognitive psychological factor of perceptual sensitivity of listeners. It was argued that differences in perceptual sensitivity of listeners to certain speech characteristics might account for different contributions of fluency aspects to fluency perception. The results from Experiments 2–4 would then mirror those from Experiment 1: breakdown and speed fluency should be well perceived but repair fluency should be perceived less accurately. RQ 2 studied the sensitivity of listeners to the three fluency aspects in three experiments that collected ratings on pausing, speed and repairs. As expected, the ratings from Experiment 2 on pausing were, of all three fluency aspects, best predicted by acoustic measures as evidenced by the highest adjusted R^2 value (Table 5). Since the subjective pause ratings were well accounted for by the objective acoustic properties of the speech, we argue that listeners are apparently most sensitive to pause characteristics of speech. Listeners are also sensitive to speed characteristics of speech, though less sensitive as compared to pause features. Surprisingly, listeners were also found to be sensitive to speech repairs. In fact, they are approximately as sensitive to speed features as they are to repairs. If perceptual sensitivity of listeners were the only factor determining the relative contributions of fluency aspects to fluency perception, then we would, based on the results from Experiment 2–4, expect to have found a larger contribution of repair measures

to the perception of fluency in Experiment 1. Apparently, listeners weigh the perceived speech characteristics on their importance for fluency judgments.

Based on the results from Experiment 1 it is evident that repair phenomena, though they are well perceived, contribute only a small amount to fluency perception. A possible account for this might be that our repair measures were not sensitive enough to expose the contribution of repair fluency to fluency perception. For instance, it has been proposed to distinguish between *error repairs* – repairing errors of linguistic form; and *appropriateness repairs* – presenting a new or rephrased message (Levelt, 1983; Kormos, 1999). Our current repair measures may have lacked the precision to adequately study the contribution of repair fluency. In addition, our repair measures only captured the frequency of occurrence of corrections and repetitions. As such, these measures are insensitive to the extent of repairs (e.g. the number of extraneous words involved). Several quick repetitions of single words may be perceived as less obstructive than lengthy garbles requiring major backtracking. However, despite the shortcomings of our repair measures, there is to our knowledge no evidence in the literature for a relation between speech repairs and fluency perception. Cucchiarini, Strik & Boves (2002) could not find any relationship between repairs and fluency perception. Repetitions also seem to differ from other types of disfluencies with respect to the online processing of speech. MacGregor, Corley & Donaldson (2009) did not find an N400 attenuation effect for repetitions or any memory effect, where these effects were established for filled pauses (Corley, MacGregor & Donaldson, 2007). Gilabert (2007) takes corrections in speech primarily as a measure of accuracy rather than fluency since corrections both denote attention to form and an attempt at being accurate. Apparently, there is no consensus on the function repairs have in speech perception. The contribution of repair phenomena to fluency perception clearly deserves more attention.

Resembling RQ 1, RQ 3 also investigated the contributions of aspects of fluency to fluency perception. Unlike previous analyses that used objective acoustic speech measurements to model subjective ratings on different perceptual dimensions, supplementary analyses were performed that used the subjective ratings on pause, speed and repair perception from Experiments 2–4 as predictors in models of the fluency ratings from Experiment 1. In this fashion, the findings from previous models could be supported or contested. These supplementary models substantiated the findings from previous models: all three aspects are involved in fluency perception but breakdown and speed fluency are most strongly related to fluency perception.

One of the limitations of the current study concerns the character of the analyses. Relationships between sets of acoustic measures and fluency perception were gauged by means of correlational analyses. One must be careful not to automatically interpret the relationships found as causal relationships (i.e. 'the fluency rating on item A was higher than item B *because of* the larger number of pauses in item B'). The present study cannot decide on the nature (e.g. direct or indirect) of the relationships that were found. Causal relationships can only be laid bare when one specific factor of interest is manipulated and all other interacting factors are kept constant (*ceteris paribus*). Future research, involving manipulating speech characteristics in different dimensions and studying its effect on fluency perception, will have to illuminate the nature

of the relationships found in the present study. Interesting in this respect would be to study effects both in L2 fluency and in L1 fluency. The current study only studied L2 fluency and therefore it remains to be shown whether pause and speed characteristics of speech also play a large role in L1 fluency perception. Based on the fact that we have shown that listeners are perceptually very sensitive to pause and speed features of speech, it may be hypothesized that a similar hierarchy of fluency aspects may be found for L1 fluency.

The fact that we have demonstrated breakdown and speed fluency to be most strongly related to fluency perception has implications for the language testing practice. With respect to automatic fluency assessment, for instance, our results indicate that speed and breakdown measures resemble human fluency perception to a very large extent. This observation corroborates the use of such measures in automatic fluency assessment. Also, from the perspective of the language learner, apparently those L2 speakers that manage to speak relatively fast with only minor pauses are more leniently judged by fluency raters than speakers who never repair at the cost of the speed of delivery and pausing. This observation may lead L2 speakers to prioritize improvements to the flow of their speech, rather than the absence of overt repairs.

Conclusion

The present study investigated the contribution of three aspects of fluency (breakdown, speed and repair fluency) to the perception of fluency. Based on comparisons between models of subjective fluency ratings, we conclude that the aspects of breakdown and speed fluency are most strongly related to fluency perception. From an investigation into the perceptual sensitivity of listeners to different speech characteristics, it was established that perceptual sensitivity is not the only factor deciding on which aspects contribute to fluency perception. Apparently, listeners weigh the importance of the perceived aspects of fluency to come to an overall judgment. This importance of fluency aspects is, then, not only determined by which speech characteristics are well perceived by the listener.

Acknowledgments

Many thanks to the researchers within the ‘What Is Speaking Proficiency’-project who kindly made their speech materials and test scores available. The WISP-project was sponsored by the Netherlands Organisation for Scientific Research, grant number 254-70-030: Margarita Steinel, Arjen Florijn, Rob Schoonen, and Jan Hulstijn, Amsterdam Center for Language and Communication, Faculty of Humanities, University of Amsterdam. We are also grateful to Theo Veenker from the UiL OTS lab for technical support, to Iris Mulders for help with participant recruitment, and to Cem Keskin and Erica Bouma for annotating the L2 speech items. Finally we would also like to express our thanks to three anonymous reviewers for their comments and suggestions.

Funding

This work was supported by Pearson Language Testing by means of a grant awarded to Nivja H. de Jong (‘Oral fluency: Production and perception’).

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*. Cambridge: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language, 59*(4), 390–412.

Bates, D., & Maechler, M. (2010). *lme4: Linear mixed-effects models using S4 classes*. Retrieved from http://CRAN.R-project.org/package=lme4. R package version 0.999375-36.

Chambers, F. (1997). What do we mean by fluency? *System*, 25(4), 535–544.

Corley, M., MacGregor, L. J., & Donaldson, D. I. (2007). It's the way that you, er, say it: Hesitations in speech affect language comprehension. *Cognition, 105*(3), 658–668.

Cucchiarini, C., Strik, H., & Boves, L. (2002). Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech. *The Journal of the Acoustical Society of America, 111*(6), 2862–2873.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (2012). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34*(1), 5–34.

De Jong, N. H., Steinel, M. P., Florijn, A. F., Schoonen, R., & Hulstijn, J. H. (in press). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language Learning, 54*(4), 655–679.

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

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics, 21*(3), 354–375.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *International Review of Applied Linguistics in Language Teaching (IRAL), 45*(3), 215–240.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications* (2nd ed.). New York: Routledge.

Hulstijn, J. H., Schoonen, R., De Jong, N. H., Steinel, M. P., & Florijn, A. F. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the common European framework of reference for languages (CEFR). *Language Teaching, 29*(2), 202–220.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics, 29*(1), 24.

Kormos, J. (1999). Monitoring and self-repair in L2. *Language Learning, 49*(2), 303–342.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System, 32*(2), 145–164.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning, 40*(3), 387–417.

Levelt, W. J. M. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 41–104.

MacGregor, L. J., Corley, M., & Donaldson, D. I. (2009). Not all disfluencies are are equal: The effects of disfluent repetitions on language comprehension. *Brain and Language, 111*(1), 36–45.

Pinget, A., Bosker, H. R., Quené, H., Sanders, T. J. M., & De Jong, N. H. (Manuscript submitted for publication). Native speakers' perceptions of fluency and accent in L2 speech.

Pinheiro, J. C., & Bates, D. M. (2000). *Mixed-effects models in S and S-PLUS.* New York: Springer Verlag.

Quené, H., & Van den Bergh, H. (2004). On multi-level modeling of data from repeated measures designs: A tutorial. *Speech Communication, 43*(1–2), 103–121.

Quené, H., & Van den Bergh, H. (2008). Examples of mixed-effects modeling with crossed random effects and with binomial data. *Journal of Memory and Language, 59*(4), 413–425.
R Development Core Team. (2010). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, www.R-project.org/
Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes, 14*(4), 423–441.
Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review/La Revue Canadienne Des Langues Vivantes, 65*(3), 395–412.
Segalowitz, N. (2010). *Cognitive bases of second language fluency*. New York: Routledge.
Skehan, P. (2003). Task-based instruction. *Language Teaching, 36*(1), 1–14.
Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510–532.
Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure and performance testing. In R. Ellis (Ed.), *Planning and task performance in a second language* (pp. 239–273). Amsterdam: John Benjamins.
Thurstone, L. L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*, 529–554.
Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of french. *Applied Linguistics, 17*(1), 84–119.
Veenker, T. J. G. (2006). *FEP: A tool for designing and running computerized experiments.* Version 2.4.19.

Appendix: Schematical representations of the scales presented to participants

In the instructions given to the participants of Experiment 1, fluency was defined as the sum of (silent and filled) pauses, speed and repairs. Judgments were given by clicking on one of the nine stars.

Experiment 1: Fluency

What is your judgment on the fluency?
not fluent at all * * * * * * * * * very fluent

Experiment 2: Breakdown fluency

What is your judgment on the use of pauses?
none and/or very short pauses * * * * * * * * * very many and/or very long pauses

Experiment 3: Speed fluency

What is your judgment on the speech rate?
very fast * * * * * * * * * very slow

Experiment 4: Repair fluency

What is your judgment on the use of repetitions and/or corrections?
no repetitions and/or corrections * * * * * * * * * very many repetitions and/or corrections