# Native speakers' perceptions of fluency and accent in L2 speech

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

*Language Testing* 2014 31: 349
DOI: 10.1177/0265532214526177

*Article*

LANGUAGE TESTING

Language Testing
2014, Vol. 31(3) 349–365

DOI: 10.1177/0265532214526177
ltj.sagepub.com
SAGE

# Native speakers' perceptions of fluency and accent in L2 speech

**Anne-France Pinget**
Utrecht University, The Netherlands

**Hans Rutger Bosker**
Max Planck Institute for Psycholinguistics, Nijmegen, The Netherlands

**Hugo Quené and Nivja H. de Jong**
Utrecht University, The Netherlands

## Abstract

Oral fluency and foreign accent distinguish L2 from L1 speech production. In language testing practices, both fluency and accent are usually assessed by raters. This study investigates what exactly native raters of fluency and accent take into account when judging L2. Our aim is to explore the relationship between objectively measured temporal, segmental and suprasegmental properties of speech on the one hand, and fluency and accent as rated by native raters on the other hand. For 90 speech fragments from Turkish and English L2 learners of Dutch, several acoustic measures of fluency and accent were calculated. In Experiment 1, 20 native speakers of Dutch rated the L2 Dutch samples on fluency. In Experiment 2, 20 different untrained native speakers of Dutch judged the L2 Dutch samples on accentedness. Regression analyses revealed, first, that acoustic measures of fluency were good predictors of fluency ratings. Second, segmental and suprasegmental measures of accent could predict some variance of accent ratings. Third, perceived fluency and perceived accent were only weakly related. In conclusion, this study shows that fluency and perceived foreign accent can be judged as separate constructs.

## Keywords

foreign accent, L2 specific fluency, native raters, perception of L2 speech, second language learners

**Corresponding author:**
Anne-France Pinget, Utrecht Institute of Linguistic OTS, Utrecht University, Trans 10, UTRECHT, 3512 JK, The Netherlands.
Email: A.C.H.Pinget@uu.nl

## Introduction

Oral fluency and foreign accent are two aspects of L2 production. They are assessed in language testing practices and influence the extent to which an L2 speaker is considered to be proficient. They are also the primary features perceived by ordinary native interlocutors, regardless of the speaker's actual proficiency (Derwing, Rossiter, Munro, & Thomson, 2004). To speak a language fluently and with a native-like accent is for many learners the ultimate goal in mastering a second language. Although the terms "fluent" and "accented" are regularly used to describe and assess someone's speech production in the L2, there seems to be no explicit consensus concerning what exactly is understood by these concepts (Chambers, 1997). The aim of this study is to advance our understanding of the concepts of fluency and accent as characteristics of L2 speech. The present study therefore investigates which variables underlie native listeners' perception of fluency and accent, and how native listeners weigh the multiple acoustic phenomena. This study has three main goals. First, we want to investigate the relationship between acoustic properties of speech and the perception of fluency by native listeners. Second, we explore the relationship between segmental and suprasegmental characteristics of speech and foreign accent as rated by native judges. Finally, we investigate the relationship between perceived fluency and perceived accent.

### *Fluency*

Lennon (1990, 2000) distinguishes between two senses of fluency. In the so-called broad sense, fluency refers to global oral proficiency (overall language performance), similar to how non-specialists tend to consider fluency. In contrast, fluency in its narrow sense is considered as one component of oral proficiency, as opposed to other components, such as grammatical knowledge or vocabulary size. Fluency is an "automatic procedural skill" (Schmidt, 1992) that encompasses the notion of "rapid, smooth, accurate, lucid, and efficient translation of thought or communicative intention into language" (Lennon, 2000, p. 26). In this study, we are concerned with fluency in this narrow sense. More recently, Segalowitz (2010, p. 48) proposed to distinguish between three facets of fluency: cognitive fluency, utterance fluency and perceived fluency. First, cognitive fluency reflects the speaker's ability to efficiently plan and execute speech by integrating the cognitive mechanisms underlying performance. Second, utterance fluency is the fluency that can be measured in a speech sample based on the acoustic properties of an utterance, such as speech rate, pausing, and repairs. Third, perceived fluency refers to the judgment that listeners make about the fluency of a speaker on the basis of their impressions drawn from the speech signal. This study focuses on the relationship between utterance fluency (objectively measurable fluency characteristics) and perceived fluency (subjective rater judgments). Our analysis of utterance fluency is based on the helpful distinction provided by Tavakoli and Skehan (2005) between three components: (1) speed fluency, that is, the speed at which speech is delivered; (2) breakdown fluency, that is, the number and length of pauses; and (3) repair fluency, that is, the number of false starts, corrections and repetitions.

Previous research on L2 fluency has mainly focused on perceived fluency. Several studies (e.g., Cucchiarini, Strik, & Bovis, 2002; Derwing et al., 2004; Kormos & Dénes, 2004; Rossiter, 2009; Ferrer, 2011) reported that pausing phenomena (i.e., breakdown fluency) and speech rate are primary factors correlating with fluency ratings. With respect to repair fluency, the literature suggests a weak relationship between repair strategies and perceived fluency. Cucchiarini et al. (2002) did not find any relationship between fluency ratings and the number of dysfluencies (which covered, among others, repetitions and corrections). Previous studies show large diversity in the L2 fluency measures that they used and, consequently, also in the outcomes that were found and conclusions that were drawn. Moreover, the issue of intercollinearity is rarely raised. The larger the number of fluency measures, the higher the probability of confounding the different measures, and therefore the higher the risk of intercollinearity if all these measures are included in a single model predicting fluency ratings (cf. Bosker, Pinget, Quené, Sanders, & De Jong, 2013). In contrast to previous studies, we investigate the correlations between acoustic measures of fluency and carefully select the orthogonal (i.e., mathematically independent) measures in order to answer the question:

RQ1: Which acoustic measures of fluency can predict perceived L2 fluency?

At first glance, it seems obvious that slower speech delivery, more or longer pauses and more repair strategies result in lower fluency judgments in L2 speech. However, studies have revealed that even native speech (L1) is not always smooth and continuous; it also exhibits hesitations and repairs (Bortfeld, Leon, Bloom, Schober, & Brennan, 2001). Moreover, a speaker who is not very fluent in his L1 (for instance, because he speaks slowly or uses a lot of pauses) cannot be expected to be very fluent in his L2. These facts suggest that researchers should take a speaker's individual L1 fluency characteristics into consideration when assessing his L2 fluency. With the exception of some recent work (e.g., Derwing, Munro, Thomson, & Rossiter, 2009; De Jong, Groenhout, Schooner, & Hulstijn, 2013), studies on L2 fluency have rarely considered the L1 fluency of their subjects. As Segalowitz (2010) noted, the fact that most researchers have not done this may have revealed individual speech differences that are in fact unrelated to the use of the L2 and have provided undesired noise that may have masked specific L2 fluency phenomena. Recently, De Jong, Groenhout, Schooner, and Hulstijn (2013) have demonstrated that a speaker's L2 fluency is indeed related to his L1 fluency. De Jong et al. reported highly significant L1–L2 correlations for different fluency measures (speech rate, pauses, etc.) and provided strong evidence that a large part of fluency-related phenomena are characteristic of the way individuals speak in general, and not just typical for their speech production in the L2. Segalowitz (2010) proposed calculating the *residuals* obtained when correlating L1 and L2 fluency measures with each other. First, the assumption is made that the most fluent speech an individual can produce is his L1 speech. L1 is therefore used as a baseline to partial out the variation which is not specifically related to the dysfluencies in L2, but which characterizes a person's general performance in the given test condition. The residualized score

expresses the difference between the actual observed value and the value predicted from the person's L1 performance. The residualized scores allow us to isolate the dysfluencies that are specifically related to the use of an L2. One goal of this study is to determine whether this new type of acoustic measures of L2 fluency proposed by Segalowitz (2010) is significantly better at explaining variance in fluency ratings than raw L2 measures:

> RQ2: To what extent do L2-specific measures (i.e., residuals) correlate with perceived fluency? Are they better predictors of perceived fluency than traditional measures of utterance fluency?

## Accent

A speaker who acquires his L2 later in life is almost certain to exhibit some degree of foreign accent (e.g., Patkowski, 1990), that is, the pronunciation of his L2 shows deviations from native norms. These deviations characterizing the speaker as a non- native are usually grouped into two categories: segmental and suprasegmental.

Segmental pronunciation errors can be phonemic or subphonemic errors. Phonemic errors are cases where a phoneme is substituted by another. In contrast, subphonemic errors are cases of allophonic variation. Suprasegmental errors are connected with suprasegmental phenomena such as stress and intonation.

Previous studies that relate the pronunciation errors in the L2 utterance to the perception of global foreign accent and its comprehensibility (Anderson-Hsieh, Johnson, & Koehler, 1992; Magen, 1998; Munro & Derwing, 1999) tend to focus on pronunciation errors that are typical for speakers of a specific L1 in the concerned L2 (e.g., Voice Onset Time (VOT) of German learners of French). Both segmental and suprasegmental factors are reported to be important for communicative effectiveness and efficiency. Segmental errors, both at phonemic and allophonic level, can hinder communication, for instance by slowing down word recognition (Smith, 2005; Munro & Derwing, 2008). Intonation, syllabic structure, lexical stress and rhythm (i.e., suprasegmental features) also help the listener to segment the speech stream and to recognize the words more quickly (Cutler & Butterfield, 1992; Cutler, 2012). Thus, segmental and suprasegmental errors both seem to make the listener perceive the speech as accented.

Three studies explicitly attempted to assess the contribution of specific phonetic and phonological factors to the perception of foreign accent. Magen (1998) investigated different segmental and suprasegmental factors in the speech of native Spanish speakers with a strong accent in English. The goal was to examine the relative weight of these different types of errors on accent ratings. Results showed that listeners were sensitive to syllable structure, final consonant deletion, consonant manner of articulation, and phrasal stress. Anderson-Hsieh, Johnson, and Koehler (1992) investigated the relationship between raters' judgments of non-native pronunciation and actual deviance in segmentals, prosody and syllable structure in 11 different languages. The accent ratings and the deviance found in each area of pronunciation showed that (1) errors in all areas have a significant influence on the ratings and (2) suprasegmental

variables proved to have the strongest influence. Kang (2010) investigated how a range of suprasegmental features independently contributed to listener's judgments of comprehensibility and accentedness. She showed that accent ratings were best predicted by pitch range and word stress measures. In conclusion, certain accent errors are related to accent ratings. Segmental and suprasegmental characteristics – typically chosen as being specific for the L2 speech of speakers with a particular L1-background – were found to be closely related to accent perceived by native listeners (e.g., Magen,1998; Anderson-Hsieh et al., 1992). In the present study, however, we focus on more global, non-language-specific segmental and suprasegmental measures of accent (e.g., goodness of segment realizations and pitch range) and their relationship to perceived accent:

RQ3: Which acoustic measures of accent can predict perceived accent?

By correlating acoustic measures with accent ratings, we want to determine how much of the variance in accent ratings can be explained by segmental and suprasegmental characteristics of the L2 speech. On the basis of previous studies, we may expect our measures to explain a non-negligible part of the variance in accent ratings as given by native listeners.

## *Relationship between fluency and accent*

Previous studies (Anderson-Hsieh & Koehler, 1988; Munro & Derwing, 1998, 2001; Derwing et al., 2004) suggest a negative correlation between accent and fluency: the stronger the perceived foreign accent, the lower the fluency ratings.

RQ4(a): To what extent are fluency and accent ratings related?

*Effects of fluency on accent ratings.* Munro and Derwing (1998, 2001) have conducted several studies that investigated whether fluency has an effect on accentedness and comprehensibility judgments. These studies systematically point to the effect of speech rate as an acoustic measure of utterance fluency on ratings of foreign accent. They tested whether increasing or decreasing the speech rate of L2 accented speech by 10% would lead to lower accent ratings. They found that fast stimuli were rated as less accented than stimuli presented at normal and slowed rates, and that the relationship between speech rate and perceived accent is curvilinear rather than linear. As long as the speed of delivery remains manageable from a processing standpoint, the listeners showed benefit from the acceleration in speech rate. However, when the same speech is presented at a particularly fast rate, the listeners may be at a disadvantage, since very fast speech places extra demands on the listener. A regression model revealed that the speaking rate could account for 15% of the variance of accent ratings.

*Effects of accent on fluency ratings.* According to a study by Freed (1995), accentedness is one of the most important factors by which raters claimed to be influenced when

reporting on their experiences during fluency rating tasks. Rossiter (2009) made a similar observation concerning the influence of pronunciation. Several studies have attempted to test this influence of accentedness on perceived fluency experimentally. However, the reported findings vary widely across studies. In her study on the role of pitch and phrasal segmentation, Wennerstrom (2000) showed that prosody affects listeners' perception of L2 fluency. Derwing and Rossiter (2003) also found that prosodic accuracy contributes to the overall impression of fluency. The underlying assumption is that inaccurate prosodic patterns are characteristic of accented speech. However, one may question whether the analysis of prosody does not directly interfere with pausing (being a component of fluency), since the prosody will necessarily be changed in a speech sample containing a large number of (too long) pauses. Therefore, it is not surprising to find a relationship between prosody and fluency in these studies. Derwing et al. (2004) examined the relationships between perceived fluency, comprehensibility and accentedness. They found a strong relationship between fluency and comprehensibility, whereas the correlation between fluency and accentedness was somewhat lower. They concluded that their findings show a relatively weak relationship between fluency and accentedness. However, the results of this study should be treated carefully, since the samples used in their study were drawn from low-proficiency speakers. It is possible that raters have judged fluency and comprehensibility more strictly than accent, since a good accent can be thought to become a requirement only when the L2 speaker reaches a higher proficiency level. In conclusion, we need more insight into the question whether accentedness is an interfering factor that plays a role when a native speaker rates L2 fluency, and vice versa:

RQ4(b): To what extent can acoustic measures of fluency predict accent ratings?
RQ4(c): To what extent can acoustic measures of accent predict fluency ratings?

To answer research questions 1–4, a *dual approach* (Cucchiarini et al., 2002) will be adopted: native perception of fluency and accent in spontaneous L2 speech will be collected in two rating experiments (the fluency ratings were collected in the same experiment as reported in Bosker et al., 2013). These ratings will be related to a number of acoustic measures of accent and fluency calculated from the speech fragments.

## Experiment 1: Fluency

### Method

*Stimuli.* Speech recordings from native and non-native speakers of Dutch were obtained from the "Unravelling second language proficiency" project from the University of Amsterdam (described in De Jong et al., 2012). Fifteen L2 speakers of both groups of the corpus (L1 English and L1 Turkish) were selected. Two L1-backgrounds were selected to allow comparison between different types of L2-speakers. The two groups were matched for proficiency on the basis of a 116 items vocabulary test, because De Jong et al. (2012) have shown that the vocabulary scores [English mean(s) = 67.5(15.7); Turkish, mean(s) = 64.1(18)] of these speakers are a good indicator of their overall proficiency. In addition to these 30 non-native speakers with an intermediate level of Dutch

proficiency, eight native speakers of Dutch were selected based on their vocabulary score being close to the native speakers' average score (mean(*SD*) = 106(5.32)). The recordings of native speakers were provided as reference points for the raters to which non-native stimuli could be compared, but they will not be further analyzed in this study.

All 38 speakers (30 L2 speakers and eight natives) performed a series of different computer-administered speaking tasks providing us with spontaneous speech in the form of conversational monologue (De Jong et al., 2012). First, the non-native speakers performed eight tasks in their L2 (Dutch). The tasks were designed for the proficiency levels B1 and B2 of the Common European Framework of Reference for Languages (Hulstijn, Schoonen, De Jong, Steinel, & Florijn, 2012). Second, the non-native speakers performed eight tasks in their mother tongue (either English or Turkish). These tasks were different, but highly similar to the ones performed in the L2.

For the analysis of the speakers' L1 fluency, recordings from all eight tasks were selected. The L1 stimuli were analyzed acoustically in order to obtain insight into a speaker's fluency behavior in his L1, but these were not presented to the raters. For the rating experiments (L2 tasks), three tasks from the eight were selected on the basis of the different task characteristics (e.g., degree of complexity, formality, discourse type). For each speech performance of approximately two minutes, 20-seconds recordings were extracted from the middle of the original recording. The stimuli were subsequently resampled to a sampling frequency of 44,100 Hz and scaled to an intensity of 70 dB. As a result, we obtained 114 speech fragments of Dutch (38 speakers × 3 tasks) for experimental use. All speech recordings were transcribed and annotated.

*Acoustic measures of fluency.* Six acoustic measures of fluency were calculated for each stimulus (see Table 1). Each measure was specific to one aspect of fluency in order to avoid confounding the three fluency aspects (speed, breakdown and repair fluency). For this reason, all frequency measures of fluency were calculated on the basis of *spoken time* (total duration excluding pauses) instead of total time (total duration including pauses). Moreover, traditionally used acoustic measures such as *speech rate, phonation time ratio* were disregarded in this study, since in such measures different aspects of fluency are confounded: they encompass information about both the speed of speech delivery and pausing patterns. Pauses shorter than 250 milliseconds were disregarded. This cut-off point of pause length is often used to separate hesitation phenomena (Towell, Hawkins, & Bazergui, 1996) from pauses that may signal the stop phase of a plosive or may be classified as micro-pauses (Riggenbach, 1991; De Jong & Bosker, 2013). For ease of interpretation, all measures share the same polarity: the higher the value, the less fluent the speech.

For speed fluency, we measured the logarithm of the mean length of syllables (MLS). The mean length of syllables is the inverse of the traditionally calculated articulation rate. The logarithm provides us with a normally distributed equivalent of the raw measure that we can use for regressions (Limpert, Stahel, & Abbt, 2001). For breakdown fluency, we selected three measures: the number of silent pauses per second spoken time (NSP), the number of filled pauses per second spoken time (NFP), and the logarithm of the mean length of silent pauses (MLP). Finally, two measures for repair fluency were used: number of corrections per second spoken time (false starts, reformulations, and

**Table 1.** List of selected acoustic measures of fluency.

| Aspect | No. | Acoustic measure | Calculation |
|---|---|---|---|
| **Speed** | 1 | Log of the mean length of syllables (MLS) | Log (spoken time/ number of syllables) |
| | 2 | Number of silent pauses per second spoken time (NSP) | Number of silent pauses /spoken time |
| **Breakdown** | 3 | Number of filled pauses per second spoken time (NFP) | Number of filled pauses / spoken time |
| | 4 | Log of the mean length of silent pauses (MLP) | Log (total length of silent pauses /number of silent pauses) |
| **Repair** | 5 | Number of corrections per second spoken (NC) | Number of corrections/ spoken time |
| | 6 | Number of repetitions per second spoken time (NR) | Number of repetitions/ spoken time |

self-corrections) (NC) and the number of repetitions per second spoken time (repetitions of exact words, syllables or phrases) (NR).

*Participants and procedure.* Twenty normal-hearing native Dutch speakers (19f/1m; mean age (*SD*) = 20.2 (1.88)) participated on a voluntary basis, and they received €7.50 for their participation. All participants were linguistically untrained native speakers with no experience in phonetics, speech therapy or rating of second language proficiency. They all came from the Randstad (a central urbanized zone in the Netherlands which comprises the major cities of Amsterdam, Rotterdam, The Hague, and Utrecht) and considered themselves as having no marked accent in Standard Dutch. Participants were seated in soundproof booths. First, written instructions were presented on the screen. The participants were told to base their judgments on the use of silent and filled pauses, the speech rate and the use of hesitations and/or corrections, and not to rate fluency in the broad sense of language proficiency. These instructions were followed by a practice phase. In the test phase, the above-described 114 speech stimuli were presented to participants using the FEP experiment software (version 2.4.19; Veenker, 2006). Participants listened to the stimuli over headphones at a comfort` volume. They rated the stimuli presented in one of six different pseudo-randomized orders using an Equal Appearing Interval Scale (Thurstone, 1928). The nine-point scale ran from 1: "not fluent at all" to 9: "very fluent". Halfway through the experiment participants were given the opportunity to pause briefly. Finally, the participants filled out a short questionnaire on their background, their attitudes towards and degree of familiarity with the speaker's foreign accent. No relationship was found between listeners' ratings and their self-rated attitude toward and familiarity with L2 speakers.

## Results

First, the speech materials were analyzed. The intercollinearity between the acoustic measures of fluency was investigated through Pearson's *r* correlations between all

measures and has already been reported in Bosker et al. (2013). The highest correlation was found between the number of silent pauses (NSP) and the mean length of syllables (MLS) ($r = .330$) Correlations between other measures were either lower or nonexistent. From this analysis, we conclude that the risk of multicollinearity in the further analyses of the results is indeed very limited.

Each item was judged by the 20 raters on fluency. The degree of agreement between these raters was very high (Cronbach's $\alpha = .97$). In order to relate the subjective ratings on each item to the acoustic measures of that item, a method of collapsing these 20 ratings for each item was required. Therefore, our analyses were performed in two consecutive steps. The first step involved correcting fluency ratings for a range of factors, such as individual differences between raters in their use of the scale, and presentation order. This resulted in corrected estimates of the raw 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 package (Bates & Maechler, 2012) in R. *Order* was added to the model as a fixed effect in order to test for general learning or fatigue effects. *Rater* (testing for individual differences between raters) was added as a random effect. *OrderWithinRaters* was added to the model as random slope on *Rater*, testing for individual differences in order effects. This most complex model was compared to less complex models (with only one or two factors). Likelihood ratio tests 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 *OrderWithinRaters* (the order effect differed among individual raters) and was used to obtain estimates of the fluency ratings. All subsequent analyses were performed on these corrected estimates, instead of on the averages of the raw ratings over the 20 raters. In a second step, acoustic measures of fluency were correlated with these estimated fluency ratings through traditional linear regression analyses.

Several linear regressions were computed to investigate to what extent the set of acoustic measures of fluency could explain the variance in fluency ratings. The adjusted proportion of variance explained ($R^2$) of these models and thus the explanatory power of each measure of fluency are presented in Table 2. First, the adjusted $R^2$ of the model with all traditional L2 fluency measures added as predictors of fluency ratings showed that 84% of the variance in fluency ratings may be explained on the basis of these six acoustic measures of fluency. Second, we computed separate models with one or more acoustic measure of fluency (within the same aspects) (see Table 2).

The discussion of these separate models and of how the different aspects of fluency load together can be found in Bosker et al. (2013). In this article, we focus on the second research question concerning the L2 specific fluency measures proposed by Segalowitz (2010): the residualized scores. We wanted to find out to what extent these residualized scores correlate with fluency ratings. We performed the same linear regressions, but now with the residualized scores instead of the L2 measures. The models were compared on the basis of the adjusted $R^2$ given in Table 2. The adjusted $R^2$ showed that the model that had all residualized measures as predictors could explain 72% of the variance. The adjusted $R^2$ of the models that had residualized measures as predictors were systematically lower than those models that had traditional L2 fluency measures as predictors (with the exception of the model with both repair fluency measures).

**Table 2.** Goodness of fit of the models of subjective fluency with measures of fluency as predictors (either L2 measures or residualized scores), expressed in adjusted R2.

| Predictors | Traditional L2 measures | Residualized scores |
|---|---|---|
| **All measures** | .837 | .717 |
| MLS | .553[a] | .545 |
| NSP | .177 | .169 |
| NFP | .010 | 0 |
| MLP | .235 | .118 |
| **NSP * NFP * MLP** | .609[a] | .416 |
| NC | .027 | .012 |
| NR | .146 | .146 |
| **NC * NR** | .175[a] | .201 |

[a]The adjusted R2 of this model differs by approximately 1% from the adjusted R2 of the same model reported in Bosker et al. (2013). This is due to a different handling of missing values in the dataset.

**Table 3.** List of selected acoustic measures of accent.

| No. | Acoustic measure | Calculation |
|---|---|---|
| 1 | Phonemic error rate (PER) | Binary Acoustic Judgment of correctness of a phoneme selection: /ə /, /ɑ/, /œy/, /ʏ/, /a/, /øː/, /ɛi/, /y/, /t/, /x/, /w/, /h/ |
| 2 | Pitch range (Pitch R) | Calculated in semitones with a reference of 60 Hz |

## Experiment 2: Accent

### *Method*

*Stimuli.* The same stimuli were used as in Experiment 1.

*Acoustic measures of accent.* Two acoustic measures of accent were calculated for each stimulus (see Table 3). The first measure is segmental in nature: the phonemic error rate (PER). Based on Neri, Cucchiarini, and Strik (2006), a study aimed at obtaining systematic information on segmental pronunciation errors made by learners of Dutch, a selection of 12 phonemes (eight monophthongs and diphthongs, and four consonants) was made that appeared to be problematic for both English and Turkish speakers. All words used in the recordings were extracted and matched with their phonetic transcription as found in CELEX (Baayen et al., 1993). In total 3512 occurrences of the selected phonemes (average 39 per speech fragment) were analyzed. For each occurrence, the first author established whether the speaker's realization was correct (i.e., matched the Dutch phoneme, coded as 1) or incorrect (i.e., when the phoneme had been substituted by another or omitted, coded as 0). Two other authors rated a smaller sample of 10% of the data. The interrater agreement between the three researchers appeared to be fair (Fleiss Kappa = .40). The *phonemic error rate* was then calculated as the proportion of incorrect

realizations of the selected phonemes divided by the total occurrence number of relevant phonemes produced in the speech sample. A low phonemic error rate (e.g., .03) represents a small number of false phonemic realizations, whereas a high phonemic rate (e.g., .42) represents a large number of false phonemic realizations.

The second measure is suprasegmental in nature: pitch range (PR). Intonation plays a large role in the cohesion of discourse, involving both the flow of information structure and the grouping of the discourse into constituents (e.g., Beckman & Venditti, 2010). A narrow pitch range might lead to the speech being perceived as monotonous. De Jong et al. (2012) showed that intonation is an important skill of language proficiency. Kang (2010) showed that pitch range was related to ratings of accentedness in English L2 data. We hypothesize that L2 speech with a narrower pitch range will be rated as having stronger foreign accent. Pitch range was measured in semitones with a reference of 60 Hz. The semitone scale is a perceptual non-linear scale of pitch that is used widely in the study of prosody and lexical tone (e.g., Liberman & Pierrehumbert, 1994; Xu, 1999).

There was no correlation between the two measures of accent ($r = -0.01$, $p = .942$). There is thus no risk of intercollinearity between these two factors.

*Participants and procedure.* To avoid ratings on one dimension (fluency) influencing the ratings on the other dimension (accent), a new group of 20 normal-hearing native Dutch speakers [17f/3m; mean age (*SD*) = 21.8 (3.85)] participated in this second experiment. Schmid and Hopp (this volume) have shown that 20 raters is a sample large enough to achieve robust data. The speech materials and the procedure of this experiment were identical to Experiment 1, but crucially the instructions given to these raters were different. The participants received precise instructions to rate accentedness, basing their judgments on the pronunciation of sounds, word stress and intonation patterns. Their ratings should represent how much the pronunciation of the speakers deviated from the norms of Standard Dutch. They rated the 114 stimuli on a nine-point scale ranging from 1: "no accent" to 9 "very strong accent".

## Results

Each item was judged by the 20 raters on accent. The degree of agreement between the raters was very high (Cronbach's $\alpha = .98$). The acoustic measures of accent were correlated with the estimates of accent ratings from Experiment 2. The goal was to investigate to what extent our two acoustic measures of accent would explain the variance in accent ratings.

In Table 4, the adjusted $R^2$ of the regressions predicting accent ratings are presented. The adjusted $R^2$ of the model with both accent measures (excluding the non-significant interaction) was .229 ($F(1;87) = 14.18$, $p < .001$) which means that 23% of the variance in accent ratings can be explained on the basis of these two acoustic measures of accent. The explained variance is relatively small, but this result is in line with previous research (e.g., Anderson-Hsieh et al., 1992; Magen, 1998). The phonemic error rate alone could explain 17% of the variance in accent ratings, and pitch range 6%.

The analysis of segmental pronunciation errors made clear that many pronunciation errors could not be captured by the PER, since a phoneme was only considered as incorrect when it was substituted by another phoneme or deleted. In that way, allophonic

**Table 4.** Goodness of fit of the models of subjective accent ratings with measures of accent as predictors.

| Predictors | Adjusted R2 |
|---|---|
| PER | .169 |
| PR | .056 |
| **PER + PR** | **.229** |

variation could not be captured (e.g., the pronunciation of /t/ as an aspirated [th] which is typical for L1 English speakers). Therefore, we computed a third measure: the phonetic quality evaluation (PQE). This measure was computed exploratively for one third of the data (= 30 items). The computation of this subphonemic measure involved (i) the segmentation and alignment of the speech material, (ii) the submission of the segmented speech material to the TQE software, and (iii) the reduction to one score per item. The Transcription Quality Evaluation (TQE) (Strik, 2012) is a tool that automatically evaluates the quality of phonetic transcriptions by determining a score for each phone (ranging from 0 to 100%) indicating the goodness of fit between the segment and its transcription. The higher the number, the better the fit. The scores were collapsed into one PQE score per stimulus, which was subsequently added to the model as predictor of accent ratings. These PQE scores did not correlate with the two other accent measures. The adjusted $R^2$ of the model with all three accent measures (PER + PQE + PR) was .339 ($F(3;24) = 5.591$, $p = .005$) which means that 34% of the variance in accent ratings could be explained on the basis of these three acoustic measures of accent on a sample of a third of the data. The computation of the *phonetic quality evaluation* allowed us to significantly improve the original model of accent ratings ($F = 6.322$, $p = .019$).

## Relating fluency and accent ratings

The fourth research question concerned the extent to which accent ratings and fluency ratings are related to each other. Combining the results from Experiment 1 and Experiment 2 is required to investigate this relationship. Accent ratings were weakly correlated with fluency ratings, ($r = -.25$, $p = .017$). The higher the accent scores, the lower the fluency scores; the stronger the perceived foreign accent, the more dysfluent the speaker is perceived to be.

In a second step, we tested whether the acoustic measures of accent might be predictors of fluency ratings. We added the two original acoustic measures of accent to the model as predictors of fluency ratings and checked whether these factors could strengthen the explanatory power. If this would be the case, one might conclude that accentedness is an interfering factor when a native speaker rates L2 fluency. However, adding two acoustic measures of accent did not result in a better fit of the model (see Table 5): Model 2 achieved an adjusted $R^2$ of .836, versus .843 for the original model ($F < 1$).

Finally, we tested whether acoustic measures of fluency may predict ratings of accent. Acoustic measures of fluency were added to the model with the two measures of accent as predictors. The question was whether these factors could add some explanation of

**Table 5.** Goodness of fit of the models of subjective accent and fluency ratings.

| Dependent variable | Predictors | Adjusted $R^2$ | Significance testing |
|---|---|---|---|
| (1) Fluency ratings | 6 fluency measures | .837 | Model 1 vs. Model 2: |
| (2) Fluency ratings | 6 fluency measures + 2 accent measures | .836 | $F(1,68) = .257$, $p = .774$ |
| (3) Accent ratings | 2 accent measures | .229 | Model 3 vs. Model 4: |
| (4) Accent ratings | 2 accent measures + 6 fluency measures | .329 | $F(1,68) = 2.188$, $p = .024$ |

variance, thus whether fluency is an interfering factor when a native speaker rates foreign accent. The basic model (Model 3) presented in Table 5 with two accent measures as predictors could account for 23% of the accent ratings. Model 4 with all acoustic measures of fluency as additional predictors reached a higher adjusted $R^2$ of .329. A comparison between Model 3 and Model 4 revealed a significant difference ($F = 2.188$, $p = .024$).

## Discussion and conclusion

This study investigated L2 oral fluency and foreign accent. In Experiment 1, 20 untrained native speakers of Dutch rated L2 Dutch speech fragments on fluency. In Experiment 2, 20 other native speakers of Dutch rated the same samples on accentedness. These ratings were subsequently correlated with a variety of acoustic measures of fluency and accent, and with each other.

Our raters appeared to be able to provide consistent and reliable judgments on L2 fluency. With regard to our first research question tested in Experiment 1, fluency ratings were predicted to a satisfactory extent by means of six acoustic predictors. Especially measures of speed and breakdown fluency could predict a large part of the variance in fluency ratings. In contrast, measures of repair fluency could explain a rather small part of the variance, but were showed to be non-negligible in contrast to what is found in previous research.

We also tested Segalowitz' proposal (2010) that residuals might be more appropriate objective measures of fluency, because the residuals partial out the role of the L1 in L2 speech. We have shown that residuals successfully predict a large proportion of the variance in fluency ratings, but they are not better than traditional L2 measures for predicting perceived fluency. After all, when listening to L2 speech from unknown speakers, raters do not have access to the L1 fluency profiles of these speakers. Therefore, taking L1 fluency into account does not result in better predictions of fluency ratings.

With respect to the third research question, investigated in Experiment 2, the segmental and suprasegmental measures of foreign accent predicted only a relatively small part of accent ratings. Both the phonetic measures and the pitch range explained some variance in accent ratings. However, the explanatory power for the accent ratings (with three predictors) remained smaller than the explanatory power achieved by the model of fluency ratings (adjusted $R^2$ of .339 vs. .837). A direct comparison between these two

models can be argued to be somewhat biased because the best model of fluency is based on all speech fragments whereas the best accent model is based on only a third of the data (30 items). However, running the model of fluency ratings with all fluency measures as predictors with a varying random sample of 30 items (the same sample size as used for the accent model), and repeating this 100 times, resulted in an adjusted $R^2$ between 0.724 and 0.905 (5% and 95% percentile limits of $R^2$, respectively) which clearly exceeds the adjusted $R^2$ of .339 for accent ratings of 30 items. We conclude that global quantitative measures of accent do not correlate as well with accent ratings as global measures of fluency do with fluency ratings. We speculate that the perception of foreign accent is also based on the qualitative analysis of a range of other factors not captured in our model (such as correct stress placement and sentence intonation).

Concerning the relationship between fluency and foreign accent (in research questions 4a, 4b and 4c), we found only a weak negative correlation between fluency ratings (from Experiment 1) and accent ratings (from Experiment 2). Speech that is rated as less fluent also tends to be rated as more accented. These results are in line with previous studies, in which accent and fluency ratings were correlated (Wennerstrom, 2000; Derwing et al., 2004). The correlation we obtained is ($r = -.25$), however, is much weaker than previously reported (Derwing et al. (2004) found $r = .49$). One possible explanation for the difference in correlation coefficients between Derwing et al. (2004) and the present study is that – in their study – the same group of raters rated accent and fluency for each stimuli at the same time. In the Derwing et al. (2004) study, listeners' judgment on the one scale might have directly influenced their judgment on the other. Another possible explanation is the amount of instructions that participants received prior to the experiment. In this study, we adopted a between group design and each group received precise instructions.

Acoustic measures of accent did not contribute to the model of perceived fluency. However, adding acoustic measures of fluency did improve the model of perceived accent. This suggests that the raters in the present study, who received detailed definitions of fluency and accent, were able to rate L2 fluency without being influenced by the accentedness of speech. This finding forces us to interpret the results of Derwing and Rossiter (2003) with caution, since their analysis of prosodic accuracy (as a measure for accent) possibly interferes with pausing (being a component of fluency). On the other hand, a speaker's L2 fluency was in the present study a predictor of the ratings of foreign accent. This finding seems to support the evidence found by Munro and Derwing (1998, 2001) and indicates that acoustic measures of fluency may play a role in perceived accent.

Having a better understanding of what makes L2 speech sound fluent or accented has valuable implications for language testing practices. The provided description of acoustic properties which contribute to the perception of dysfluency and accentedness might help individual raters to provide a more objective and reliable assessment of the L2 linguistic performance. Moreover, software for automatic assessment of fluency might be improved in view of the proposed shortlist of powerful acoustic measures, in order to better reflect native ratings of fluency. In large-scale language assessments, where fluency and accent rated by human judges are seen as indicators of L2 proficiency, these two aspects are not always kept separated. The TOEFL test, for instance, has a rubric to

rate the speaking tasks called "delivery", which combines aspects of fluency and pronunciation. Our results, however, suggest that human raters are capable of rating these two aspects separately, and crucially a speaker who is not very fluent, does not necessary have a strong foreign accent, and vice versa. Ratings on "delivery", covering both fluency and accent, may thus prove problematic for raters since the two concepts can very well exist independently from the other.

In conclusion, this present study has shown that native speakers' perception of L2 fluency is highly correlated to the quantitative, temporal properties of L2 speech, whereas native speakers' judgments of foreign accent is only weakly dependent on acoustic accentedness measures. A more qualitative and probably L2 specific analysis of accent features is required to understand native speakers' perception of foreign accent.

## Acknowledgments

We thank the researchers of the "What Is Speaking Proficiency"-project (WISP) who kindly made their speech materials and test scores available for this study. 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 very grateful to René Kager for his useful comments, to Theo Veenker and Iris Mulders from the UiL OTS lab for technical support and for participant recruitment, and to Cem Keskin and Erica Bouma for annotating the speech material. We would like to thank the anonymous reviewers for their insightful comments on the earlier version of this paper.

## Funding

This work was supported by Pearson Language Testing by means of a grant awarded to the second and fourth author ("Oral fluency: production and perception").

## References

Anderson-Hsieh, J., Johnson, R., & Koehler, K. (1992). The relationship between native speaker judgments of nonnative pronunciation and deviance in segmentals, prosody, and syllable structure. *Language Learning*, *42*, 529–555.

Anderson-Hsieh, J., & Koehler, K. (1988). The effect of foreign accent and speaking rate on native speaker comprehension. *Language Learning*, *38*, 561–613.

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.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The CELEX lexical database* (CD-ROM). Linguistic Data Consortium, University of Pennsylvania, PA.

Bates, D., & Maechler, M. (2012). *lme4: Linear mixed-effects models using S4 classes*. Available: http://CRAN.R-project.org/package = lme4. R package version 0.999999–0.

Beckman, M. E., & Venditti, J. J. (2010). Tone and Intonation. In W. J. Hardcastle & J. Laver (Eds.), *The handbook of phonetic sciences* (pp. 603–652). Oxford: Blackwell.

Bortfeld, H., Leon, S., Bloom, J., Schober, M., & Brennan, S. (2001). Disfluency rates in conversation: Effects of age, relationship, topic, role, and gender. *Language and Speech*, *44*, 123–147.

Bosker, H. R., Pinget, A-F., Quené, H., Sanders, T. J. M., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing 30*(2), 159–175.

Chambers, F. (1997). What do we mean by fluency? *System*, *25*, 535–544.
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.
Cutler, A. (2012). *Native listening: Language experience and the recognition of spoken words*, Cambridge, MA: MIT Press.
Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, *31*, 218–236.
De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Ed.), *Proceedings of disfluency in spontaneous speech* (pp. 17–20). DiSS 2013.
De Jong, N. H., Groenhout, R., Schooner, R., & Hulstijn, J. H. (2013). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*; doi: 10.1017/S0142716413000210.
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*, 5–34.
Derwing, T., Munro, M., Thomson, R., & Rossiter, M. (2009). The relationship between L1 fluency and L2 fluency development. *Studies in Second Language Acquisition*, *31*(4), 533–557.
Derwing, T., & Rossiter, M. (2003). The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech. *Applied Language Learning*, *13*, 1–18.
Derwing, T., Rossiter, M., Munro, M., & Thomson, R. (2004). L2 fluency: Judgments on different tasks. *Language Learning*, *54*, 655–679.
Ferrer, M. (2011). The development of oral fluency and rhythm during a study abroad period. Doctoral dissertation. University of Pompeu Fabra, Barcelona, Spain.
Freed, B. (1995). What makes us think that students who study abroad become fluent? In B. Freed (Ed.), *Second language acquisition in a study abroad context* (pp. 123–148). Amsterdam: John Benjamins.
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 Testing*, *29*, 203–221.
Kang, O. (2010). Relative salience of suprasegmental features on judgments of L2 comprehensibility and accentedness. *System*, *38*, 301–315.
Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, *32*, 145–164.
Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language Learning*, *40*, 387–412.
Lennon, P. (2000). The lexical element in spoken second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 25–42). Ann Arbor, MI: The University of Michigan Press.
Liberman, M., & Pierrehumbert, J. (1984). Intonational invariance under changes in pitch range and length. In M. Aronoff & R. Oehrle (Eds.), *Language sound structure* (pp. 157–233). Cambridge, MA: MIT Press.
Limpert, E., Stahel, W., & Abbt, M. (2001). Lognormal distributions across the sciences: Keys and clues. *BioScience*, *51*(5), 341–352.
Magen, I. (1998). The perception of foreign-accented speech. *Journal of Phonetic*s, *26*, 381–400.
Moyer, A. (1999). Ultimate attainment in L2 phonology: The critical factors of age, motivation and instruction. *Studies in Second Language Acquisition*, *21*, 81–108.
Munro, M. J., & Derwing, T. M. (1998). The effects of speech rate on the comprehensibility of native and foreign accented speech. *Language Learning*, *48*, 159–182.

Munro, M. J., & Derwing, T. M. (1999). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*, *49*, Supplement 1, 285–310.

Munro, M. J., & Derwing, T. M. (2001). Modeling perceptions of the accentedness and comprehensibility of L2 speech: Accentedness and comprehensibility of L2 speech: The role of speaking rate. *Studies in Second Language Acquisition*, *23*, 451–468.

Munro, M. J., & Derwing, T. M. (2008). Segmental acquisition in adult ESL learners: A longitudinal study of vowel production. *Language Learning*, *58*, 479–502.

Neri, A., Cucchiarini, C., & Strik, H. (2006). Selecting segmental errors in L2 Dutch for optimal pronunciation training. *IRAL – International Review of Applied Linguistics*, *44*, 357–404.

Patkowski, M. (1990). Age and accent in a second language: A reply to James Emil Flege. *Applied Linguistic, II*, 73–89.

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*, 413–425.

Riggenbach, H. (1991). Towards an understanding of fluency: A microanalysis of nonnative speaker conversation. *Discourse Processes*, *14*, 423–441.

Rossiter, M. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *The Canadian Modern Language Review*, *65*(3), 395–412.

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

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. London/New York: Routledge.

Smith, R. (2005). The role of fine phonetic detail in word segmentation. Doctoral dissertation. University of Cambridge, UK.

Strik, H. (2012).Transcription evaluation quality tool. Retrieved from http://vps8639.xlshosting.net/TQE/

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*, 84–119.

Veenker, T. J. G. (2006). *FEP: A tool for designing and running computerized experiments* (version 2.4.19) [Computer program].

Wennerstrom, A. (2000). The role of intonation in second language fluency. In H. Riggenbach (Ed.), *Perspectives on fluency* (pp. 102–127). Ann Arbor, MI: The University of Michigan Press.

Xu, Y. (1999). Effects of tone and focus on the formation and alignment of F0 contours. *Journal of Phonetics*, *27*, 55–105.