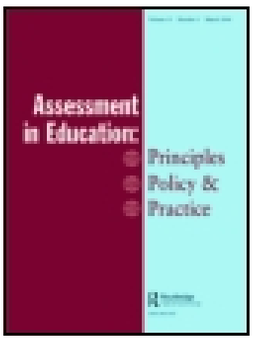

Routledge
Taylor & Francis Group

Assessment in Education: Principles, Policy & Practice

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/caie20

PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically

Nivja H. de Jong, Jos Pacilly & Willemijn Heeren

To cite this article: Nivja H. de Jong, Jos Pacilly & Willemijn Heeren (2021): PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically, Assessment in Education: Principles, Policy & Practice, DOI: 10.1080/0969594X.2021.1951162

To link to this article: https://doi.org/10.1080/0969594X.2021.1951162

ASSESSMENT IN EDUCATION: PRINCIPLES, POLICY & PRACTICE
https://doi.org/10.1080/0969594X.2021.1951162

Routledge
Taylor & Francis Group

PRAAT scripts to measure speed fluency and breakdown fluency in speech automatically

Nivja H. de Jong [a,b], Jos Pacilly[a] and Willemijn Heeren[a]

[a]Leiden University Centre for Linguistics (LUCL, Leiden University, Leiden, The Netherlands; [b]ICLON Graduate School of Teaching, Leiden University, Leiden, The Netherlands

ABSTRACT

Fluency in terms of speed of speech and (lack of) hesitations such as silent and filled pauses ('uhm's) is part of oral proficiency. Language assessment rubrics therefore include aspects of fluency. Measuring fluency, however, is highly time-consuming because of the manual labour involved. The current paper aims to automatically measure aspects of L2 fluency, including filled pauses, in both Dutch and English. A revised existing script and a new script for filled pauses are tested on accuracy. We also gauged whether the outcomes of the new script could be used for language assessment purposes by relating the outcomes to human judgements. Without further investigations, the current script should not (yet) be used for the purpose of assessing fluency automatically in (high-stakes) oral proficiency assessment. However, the performance of the scripts for measuring aspects of fluency globally and quickly are promising, especially given their stability in accuracy on new corpora.

ARTICLE HISTORY
Received 2 July 2020
Accepted 26 June 2021

KEYWORDS
Second language speech; fluency; filled pauses; PRAAT-script

Introduction

Speaking is a remarkable skill. Before a speaker is able to articulate the appropriate sounds, a number of speech production processes have been carried out and have been carried out quickly. There are roughly three stages in speech production: conceptualising what to say, formulating how to say this in language, and finally articulating the appropriate sounds (e.g. Levelt et al., 1999). If at any stage of the speech production process the speaker encounters a problem, the speaker will become disfluent, which may result in silent pauses, filled pauses (e.g. 'uh', 'uhm'), or slowing down articulation speed.

When speaking in a second language (L2), the same stages are needed to proceed from thoughts to articulated sounds. However, it will be more difficult, because the processes are less automatised in the L2, especially those needed for linguistic formulation of the message. Moreover, the L2 linguistic knowledge that is needed may at times be insufficient, which may cause disfluencies, for instance, when the speaker decides to reconceptualise, circumventing the need for the specific L2 linguistic knowledge (see Segalowitz, 2010 for an elaborate description of disfluencies in L2 speaking). In short, disfluencies in speech are telling of speaking proficiency: only highly proficient L2 learners with highly

CONTACT Nivja H. de Jong n.h.de.jong@hum.leidenuniv.nl Leiden University Centre for Linguistics (LUCL), Leiden University, The Netherlands

Supplemental data for this article can be accessed here.

developed L2 knowledge and skills will be able to fluently express their thoughts, without undue hesitations. In addition to this theoretic reasoning, there is ample research showing that holistic evaluations of proficiency are (highly) related to measures of fluency in speech (Ginther et al., 2010; Iwashita et al., 2008; Kahng, 2014; Kang et al., 2010; Révész et al., 2016). It is therefore no wonder that language tests include aspects of fluency in their rubrics to distinguish levels of proficiency as in the often-used Test of English as a Foreign Language internet Based Test (TOEFL iBT), the International English Language Testing System Academic (IELTS), the oral proficiency interview of the American Council on the Teaching of Foreign Languages (ACTFL OPI), and in the Pearson Test of English Academic (PTEA).

Having established the importance of fluency as part of oral proficiency, we note that measuring aspects of fluency are highly time-consuming. To aid future research into the relation between specific aspects of fluency and proficiency, and potentially for the purpose of assessing fluency automatically in language testing, the present study set out to investigate to what extent aspects of fluency may be evaluated automatically.

Theoretical background

De Jong and Wempe (2009) developed a tool that measures some aspects of fluency: silent pauses (frequency and duration) and speed of speaking, automatically, without the need for transcribing or even manual annotations and measurements. Although the script has been used to measure these aspects of pausing and speed, information on filled pauses such as 'uhm' and 'uh' is, as yet, missing, although filled pauses have been shown to distinguish between L1 and L2 speech (e.g. De Jong, 2016; Kahng, 2014), to be strong predictors of communicative adequacy in task fulfilment (Révész et al., 2016), and to be related to overall L2 proficiency (De Jong, 2016).

The present study, rather than using automatic speech recognition, will only aim to detect filled pauses. It is (as yet) not feasible to arrive at correctly recognised speech including filled pauses for L2 speakers with many diverse accents, in multiple languages (but see recent developments for English, Moussalli & Cardoso, 2020). Instead, this study will use general, acoustic properties of the signal to detect filled pauses. The next paragraph will describe previous research that has indicated acoustic characteristics of filled pauses so far. In short, these include duration in milliseconds, overall pitch, and vowel qualities. Overall pitch can be measured through F0, which is the fundamental frequency of a sound, resulting from the rate of vibration of the vocal folds. With respect to the vowel qualities, F1 and F2 are the main formants that shape vowels, resulting from changes in the shape of the vocal tract (e.g. by changing the shape and location of the tongue). F3 is the third resonance in the speech signal, likewise resulting from changes in the shape of the vocal tract (e.g. by lip-rounding).

A number of studies (Hughes et al., 2016a, 2016b; Kaushik et al., 2010; Shriberg, 2001, p. 165; Stouten & Martens, 2003, p. 3) have indicated duration of the syllable or vowel within the syllable to be a significant characteristic of filled pauses. The other characteristics indicated so far are variation of F0 (Verkhodanova & Shapranov, 2016), the height of F0 (Clark & Fox Tree, 2002; Shriberg & Lickley, 1993), variability in formants F1 through F3 (Audhkhasi et al., 2009; Kaushik et al., 2010), and overall stability (Stouten & Martens, 2003, p. 4). From these studies it can be concluded that filled pauses, in contrast

to other syllables, tend to have longer durations, show less F0-variation, have a lower F0, and less F1-F3 variability. In other words, filled pauses tend to be long, stable or steady syllables pronounced at a low pitch. Additionally, filled pauses are usually pronounced as 'lazy' vowels or close to a schwa ([ə]) (Shriberg, 2001). For (American) English, Vasilescu and Adda-Decker (2007) have shown that the sound of the filled pause may be closer to a mid-open back unrounded vowel ([ʌ]). For back vowels such as /ʌ/, it is known that the F1 and F2 are relatively close to each other (Reetz & Jongman, 2009, p. 184). For both such a back unrounded vowel as well as for a schwa-sound, the F3 can be relatively high, with lips not being rounded (Ladefoged & Johnson, 2011). Another potential variable that may distinguish between 'lazy' (thus filled pauses) and non-lazy syllables would be a variable that captures to what extent the speaker makes an effort in their articulation, hence to what extent the current vowel is further away from what can be said to be the default or average vowel of that speaker. For this, we propose to include variables that measure the distances between F1, F2, F3 to the relative medians of the F1, F2, F3 of that speaker.

The studies that have created algorithms to detect filled pauses are difficult to compare to each other and difficult to evaluate for the purpose of measuring L2 filled pauses. Firstly, they are difficult to compare because they differ in the (combinations of) features being tested. Likewise, different kinds of learning algorithms have been used (e.g. Gaussian mixture models as in Krikke & Truong, 2013, or no learning phase as in Kaushik et al., 2010). Also, methods differed in what time-unit was used during the training and testing phase of the algorithm; for instance, a window of 110 milliseconds (Audhkhasi et al., 2009), very short windows (e.g. Verkhodanova & Shapranov, 2016), or Stouten et al. (2006) who first divided the speech into silent and phoneme-like segments. Finally, none of the studies reported automatic filled pause detection on speech by L2 speakers.

Based on the current body of knowledge on acoustic properties of filled pauses briefly described above, the present research attempted to arrive at an easy-to-use implementation of a filled pause detection algorithm. Most of the studies above have investigated English filled pauses. Stouten and Martens (2003) propose that their method for detecting Dutch filled pauses would work for English as well. The present study aimed to implement algorithms for both English L2 and Dutch L2 filled pauses. We chose Dutch in addition to English because filled pauses in Dutch and English seem to have similar acoustic properties. The script by De Jong and Wempe (2009) was written in PRAAT (Boersma & Weenink, 2016), a computer program that enables precise analyses of speech, and includes a scripting language. We will use the same program to test and implement the algorithms in the current research.

Research aims

Building on the previous research described above, the research aims of the present study were as follows:

(1) To create a PRAAT script that measures aspects of L2 fluency automatically, including information on silent pauses, filled pauses, and speed of speech. This script will run without the need to transcribe or recognise speech, and without the

need to have separate (annotated) speech data available for training of the specific algorithm;
(2) To test the accuracy of the script with respect to the detection of filled pauses for two types of speech data (Dutch and English-speaking performances in language assessment settings), by relating the outcomes of the script to manual annotations of filled pauses;
(3) To gauge validity of the automatic measures of fluency for the purpose of language assessment, by relating the outcomes of the script to judgements on fluency.

Methods

For the purpose of training and testing an algorithm that automatically measures fluency in L2 speech, two existing corpora of Dutch and English L2 speech were chosen as primary materials. In the Dutch corpus, the performances were elicited for research into assessment and in the English corpus, the materials were actual assessment performances. Choosing existing L2 assessment corpora has the advantage that the data are ecologically valid. At the same time, choosing these corpora meant that sound quality was lower than is optimal for precise phonetic measurements in the case of the English corpus, and that available sound files were relatively short (20 seconds) to allow for ideally stable measurements in the case of the Dutch corpus. However, we also added secondary corpora unrelated to language assessment to test the resulting algorithms for more general use. Finally, necessary for the third research aim, judgements on fluency were already available for both primary corpora. In what follows, we describe the corpora (materials) in some detail and the methods used in creating the PRAAT-scripts.

Materials

English primary corpus (APTIS)

The English corpus consisted of a subset of the speech data (120 items with a total duration of about 240 minutes) as described in more detail in Tavakoli et al. (2017), including annotations and human judgements. This corpus, including annotations and judgements, was made available to us by the British Council. The corpus consists of performances from the operational APTIS speaking test, which is a computer-based speaking test consisting of four tasks, taking about 12 minutes to complete. Test takers in this corpus had a range of L1 backgrounds (Arabic, Bengali, Bosnian, French, Georgian, German, Greek, Japanese, Malay, Mandarin, Romanian, Spanish, Tamil, Ukrainian, and Uzbek – as guessed from the location of the test takers). Performances on the APTIS tasks 1 through 4 were included, targeting the A1/A2 level (task 1) through B2 (task 4). Tasks 1 through 3 consist of three prompts each, and test-takers have 30 seconds (Task 1) or 45 seconds (Tasks 2 and 3) to respond to each prompt. For task 4, participants get 1 minute of preparation time, and have a maximum time of 2 minutes to respond. Tavakoli et al. (2017) selected eight speakers from each level from A2 through C1.[1] Task 4 of the A2-level test-takers was not included in their study, since these recordings rarely contained enough speech. The subset (60 files) for the current study was selected by one of the annotators (see below) on sound quality. Note, however, that the sound quality of the English files in general was below normal standards for phonetic analyses.

APTIS speaking performances are rated holistically. Analytical descriptors, such as descriptors for fluency, are included. The performances selected by Tavakoli et al. (2017) were always rated with the same score holistically as analytically on fluency. The fluency descriptors mention pausing, false starts, and reformulations at levels A1 through C1, where the A1 descriptor mentions such disfluencies impede understanding and the C1 descriptor mentions such disfluencies do not interrupt the flow of speech (see Table 3 in Tavakoli et al., 2017, p. 12).

As described in Tavakoli et al. (2017) human annotations, including annotations on timings, were available. More specifically, the speech data were transcribed and using PRAAT, exact locations and timings of both silent and filled pauses were annotated by Tavakoli et al. (2017). Minimum (silent or combined with filled) pause time was set to 250 ms.

Dutch primary corpus

The Dutch primary corpus consisted of speech data (114 items of 20 seconds each), annotations, and human judgements as described in more detail in Bosker et al. (2013). This corpus includes 15 English, 15 Turkish, and eight native speakers of Dutch performing three tasks targeted at the B1 and B2 level (see Hulstijn et al., 2012). These tasks were monologues (max two minutes) in which participants were engaged in a long turn where context and task was described, which is comparable mostly to task 4 of the APTIS tasks (see primary English corpus below). The speech materials thus consisted of 38 speakers performing three tasks (= 114 items). Fragments of approximately 20 seconds were excerpted from around the middle of the original recordings. Each fragment started at a phrase boundary (Analysis of Speech Unit; Foster et al., 2000) and ended at a silent pause (>250 ms).

In the Bosker et al. (2013) study, judgements on fluency were made by 20 listeners who received specific instructions to rate on 'fluency': the use of silent and filled pauses, the speed of delivery of the speech, and the use of hesitations and/or corrections (and not on grammar, for example). Participants rated the speech fragments using a nine-point scale.

Bosker et al. (2013) calculated objective acoustic measures for the 90 L2 recordings, based on human transcripts of the speech recordings, and using PRAAT to measure silent pausing measures (threshold of silent pauses was 250 ms). Although these transcripts included information on filled pauses, for the purpose of the current study, the exact location and timing of these filled pauses still had to be added.

Secondary corpus

As for the secondary corpus, a subset of the D-LUCEA corpus (Orr & Quené, 2017) was used. This subset consisted of L1 Dutch and L2 English speech data from the same 59 female speakers, and L1 English (mixed American and British) from 12 male and female speakers. All speakers performed multiple speaking tasks, but here a two-minute informal monologue was used. This was produced in the speakers' L1, and if that was not English, also in their L2. L2 speakers are estimated to at least be at the B2 level (see De Boer & Heeren, 2020). For these recordings no fluency judgements were available.

Annotations of the corpora

We used PRAAT to manually annotate, in a PRAAT TextGrid (PRAAT-compatible annotation tool), the beginning and ending of each filled pause, adding precise timings

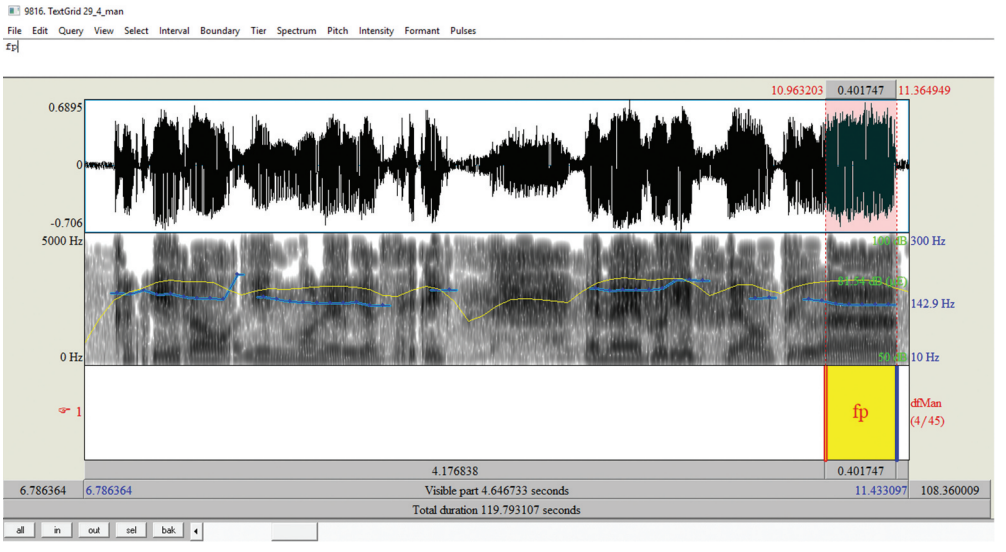

Figure 1. PRAAT screenshot of visualisation of sound file with corresponding TextGrid as created by the annotator; as can be seen, a filled pause ('fp') has manually been annotated, starting at 10.963 and ending at 11.365 seconds.

Table 1. Total number of syllables and filled pauses (as indicated by at least one annotator) in the different corpora.

	Total syllables	Filled pauses (manual)
L1 Dutch primary corpus	6,651	577
L2 English primary corpus	19,808	1,585
L1 Dutch secondary corpus	24,006	1,135
L2 English secondary corpus	22,366	1,105
L1 English secondary corpus	4,179	193

of filled pauses. Such annotations were not available for the Dutch primary corpus, nor for the English L1 speakers in the secondary corpus. Thus, for each sound file, and by each annotator separately, a TextGrid was created with annotations of filled pauses ('fp') and their timings. See Figure 1 for a screenshot of the visualisation of a soundfile with the corresponding TextGrid. Two annotators annotated the primary corpora. For the secondary corpus, data from the Dutch L1 speakers, in both languages, had already been annotated by at least two coders (see De Boer & Heeren, 2020). For the English L1 speakers in the secondary corpus, filled pauses were annotated by one coder (3rd author). Table 1 shows the total numbers of syllables, as well as the number of manually annotated filled pauses (by at least one annotator) in all corpora and subsets.

PRAAT scripts

New version of script to detect syllables

The existing script by De Jong and Wempe (2009) to detect syllables was rewritten. The original script first of all distinguishes between silences and sound (assumed to be

speech) by applying an intensity threshold (in dB), and then detects syllable nuclei. These are detected as voiced peaks in intensity, preceding and followed by dips in intensity. We now adapted the scripting syntax to the latest PRAAT programming language. We also made some changes in efficiency without changing the original way to detect syllable nuclei but small differences in millisecond timings of syllables are likely to occur (for instance, because sound and silence are detected in a slightly different manner). In Appendix 1, we describe the steps of the rewritten script to detect syllables in more detail. The new script itself can be found in Appendix 2 (supplemental appendices can be found online).

Measures to detect filled pauses

From the literature on filled pauses, an initial PRAAT script was created that measured, for each syllable as indicated by the syllable-detection script, duration, properties of the F0, and of the formants F1, F2, and F3. The script takes the TextGrid with syllable peaks (output of new syllable-nuclei-script) and the corresponding sound file as input, and takes the following steps:

(1) Syllable boundaries are detected as 6 dB in intensity below the syllable peak at both sides of the peak. When this leads to no boundary between syllables (i.e. when the dip between the syllables is less than 6 dB but more than 2 dB), the boundaries for the two adjacent syllables are put at the lowest intensity, set at 0.0001 ms apart. Syllables are then put as intervals in the TextGrid and held in an array;
(2) All syllables as detected in step 1 are concatenated into one stretch of sound (with 10 ms overlap);
(3) The median F0 of the concatenated sound is calculated to obtain a global estimate of the speaker's (sound file) voice; This variable is used to determine pitch ceilings and maximum formant values;
(4) The median F0 of each syllable is measured and put in an array, using a pitch ceiling of 2.5 * global median F0.[2] (NB: undefined values for F0 are first replaced by the mean);
(5) Median global F1, F2, and F3 are measured, using a maximum formant value of 4000 + 4 * (global median F0 – 100)[3];

Using these settings for pitch ceiling and maximum formant values, the following variables were measured for each syllable:

- Duration in seconds;
- Duration, z-normalised; i.e. the duration of each syllable, relative to the mean duration of all syllables of that speaker (of that sound file), in standard deviations from that mean;
- F0z: fundamental frequency in semitones (reference value 100 Hz), z-normalised. Thus, −1 would mean 1 standard deviation below the speaker's median F0;
- sdF0: standard deviation of the F0 in semitones;
- Distance between F1 and F2 (in Bark);
- F3 (in Bark);

- Standard deviations of F1, F2, F3 (in Bark);
- Mean absolute deviations of F1, F2, F3 to the globally measured F1, F2, F3 (of that speaker/sound file, in Bark);

Based on the theoretical background, we expect filled pauses, in comparison to other syllables, to have longer absolute and/or z-normalised durations, lower pitch (F0z), lower standard deviations of F0 and of F1 through F3, a lower distance between F1 and F2, a higher F3, and lower deviations of F1 through F3 to their respective global measures of F1 through F3.

Defining the algorithms to detect filled pauses

Data preparation. First, because exact beginnings and endings of syllable boundaries and thus of filled pauses cannot be defined or perceived precisely and are therefore somewhat arbitrary in manual annotations, we allowed for some mismatch in the timings of automatic measurements of syllable boundaries to those of the manual measurements (of filled pauses). For both annotators (Dutch data), and for all three annotators (English data) of the primary corpora, we matched manually detected filled pauses to the automatically defined syllables when the manual annotated filled pause fell between 120 ms before and/or 120 ms after an automatically defined syllable.

Secondly, we combined the annotations by the annotators with the outcomes of all measures for all automatically detected syllables. If one of the annotators had annotated a certain automatically defined syllable to be a filled pause, we took this syllable as a filled pause. In other words, the algorithm was trained to detect all manually detected filled pauses (including those when only one of the annotators had noted the filled pause).

Thirdly, we deleted syllables with extreme values for one of the measures. This led to deletion of 0.53% (35/6,651) of the syllables in the Dutch corpus and 0.12% (23/19,808) of the syllables in the English corpus.

Defining algorithms. In Figure 2, please find a scheme in which the steps following data preparation are shown. Firstly, the data in both corpora was split into a training set and a test set. The training set consisted of 70% randomly selected syllables in both corpora. We then adopted a repeated cross-validation approach to find an algorithm to detect filled pauses for each corpus, using this training set. Within the cross-validation, we used generalised linear modelling predicting the binomial variable 'FP' or 'normal' (not FP), as determined by manual annotations. For each step in the analyses, the cross-validation was carried out with 10 folds (James et al., 2013, p. 184). This means that for each out of 10 times that we applied a model, a different 10% of the data is kept as a test-set. The squared error (predicted versus actual response) on this test set is saved and then averaged over the 10 folds. This averaged squared error (delta) for a particular model was saved. The cross-validation for one model was repeated 10 times, resulting in 10 delta-values for one generalised linear model predicting the binomial variable. The 10 deltas were then compared to another set of 10 deltas from a competing algorithm, using a t-test (alpha was set to 0.01). A significant difference in which delta's from a more complex model were lower than those from a more parsimonious model would mean that the more complex model outperforms the simpler model and thus the more complex model is kept. In this way, a number of variables were found not to significantly improve the algorithm to detect filled pauses and were deleted from the

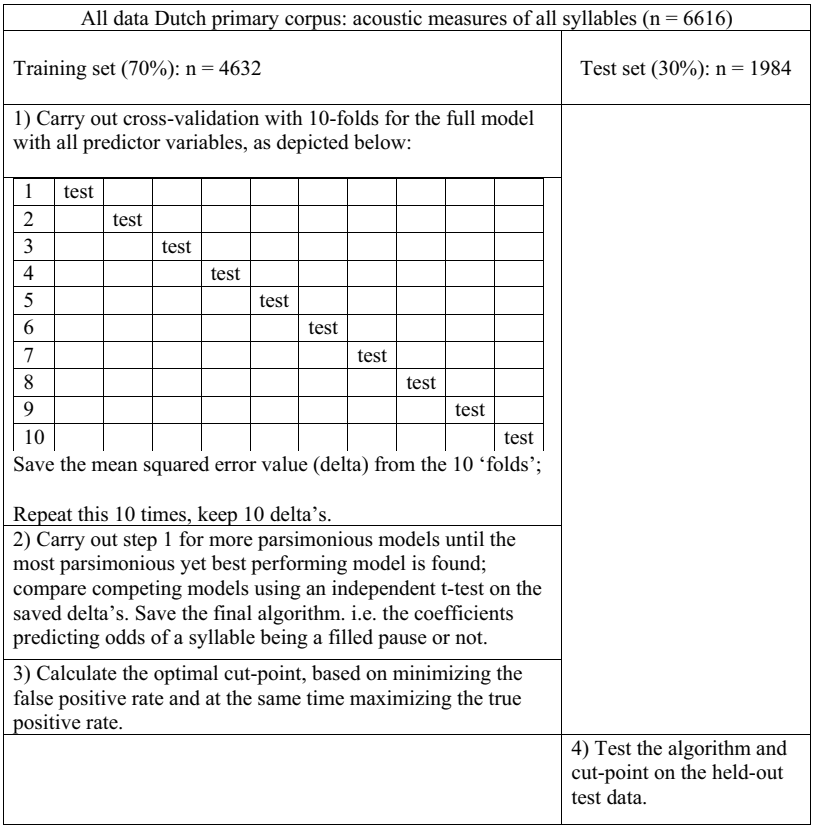

Figure 2. Scheme depicting steps in defining the algorithm and cut-point, and testing the algorithm (Dutch primary corpus as example).

algorithm. Additionally, we deleted variables which led to improved algorithms (as evidenced by the difference in the t-test) but were inconsistent with theory. For the English data, for example, contrary to visual inspection of the histograms for the annotated filled pauses versus all other syllables, and contrary to theory, higher values of the (square root transformed) standard deviation of the F3 would lead to a higher probability of a filled pause. Most likely, this finding is due to collinearity in the data (standard deviations of F1-F3 correlate with each other, plus they correlate with the mean absolute deviations of F1-F3). With high collinearity, spurious findings are known to occur (Belsley, 1991). By testing subsequent models in this way (beginning with the full model and leaving out non-significant and illogical predictors) in the training set, we derived two formulae, one for each corpus.

For the Dutch corpus, the formula according to the generalised linear model on the training set was:

Score = 8.62× sqrt(duration) − 0.36× F0z − 0.11×(F2 − F1) + 0.21× F3 − 1.36× sqrt (standard deviation of F2) − 1.02× sqrt(standard deviation of F3) − 0.72× sqrt(absolute deviation of F1) − 1.62× sqrt(absolute deviation of F2)

And for the English corpus, the formula was:

Score = 4.73× sqrt(duration) − 0.29× F0z − 0.20×(F2 − F1) + 0.31× F3 − 0.32× sqrt (standard deviation of F1) − 1.38× sqrt(standard deviation of F2) − 0.10× sqrt(absolute deviation of F1) − 0.80× sqrt(absolute deviation of F2)

The difference between the formulae is that for the Dutch corpus (and not the English corpus), the standard deviation of F3 was kept in the model, and that for the English corpus (and not for the Dutch corpus) the standard deviation of F1 was kept in the model. For the other variables, the difference is found in the coefficients. For instance, for the Dutch data, the (square root-transformed) duration variable counts almost twice as much compared to the English data. Contrastingly, the F2− F1 coefficient is twice as high for the English data compared to the Dutch data.

Defining optimal cut-points. Cut scores were determined by using the *cutpointr* package (Thiele et al., 2019) in R. To choose the optimal cut point, the procedure calculates, for each potential cut point, the true positive rate and the false positive rate. Consider the so-called confusion matrix in Table 2. The true positive rate is the number of true positives divided by all manually annotated filled pauses, so true positives + false negatives. The false positive rate is the number of false positives divided by all manually annotated normal syllables, so all false positives + true negatives. The scores for all syllables resulting from the algorithm in the training set of the Dutch primary corpus ranged from −3.893 to 7.188. If the cut score would be set to the minimum score (−.3.893), all syllables above this score would be classified as filled pauses, hence all syllables would be classified as filled pauses. This would lead to a true positive rate of 1, meaning that all actual filled pauses are correctly classified, but at the same time it would also lead to a false positive rate of 1, meaning that all normal syllables are wrongly classified as being a filled pause. If the cut point is set higher, the false positive rate will go down (which is good, fewer normal syllables wrongly classified as filled pauses), but at the same time, the true positive rate will also go down (which is a shame, fewer filled pauses correctly classified as such). The Receiver Operating Characteristic (ROC) curve depicts this relationship between the false positive rate on the x-axis and the true positive rate on the y-axis for all thresholds. Figure 4 shows the ROC-curve for the Dutch training data. Figure 3 shows the distributions of all manually filled pauses (left panel) and syllables not annotated as filled pauses (normal syllables, right panel) with the chosen cut point, as found by the optimisation procedure.

For the Dutch corpus, the optimal cut-point was found to be 2.7094, as shown in Figure 4. Subsequently, all syllables with scores higher than 2.7094 in the Dutch test set were then annotated as automatically-detected filled pauses. For the English corpus, the optimal cut-point was found to be 3.4942. Thus, all syllables with scores higher than 3.4942 in the English test set were annotated as automatically-detected filled pauses. The

Table 2. Illustration of a confusion matrix.

		Manual annotation	
Automatic detection		FP	'normal'
	FP	True positive	False positive
	'normal'	False negative	True negative

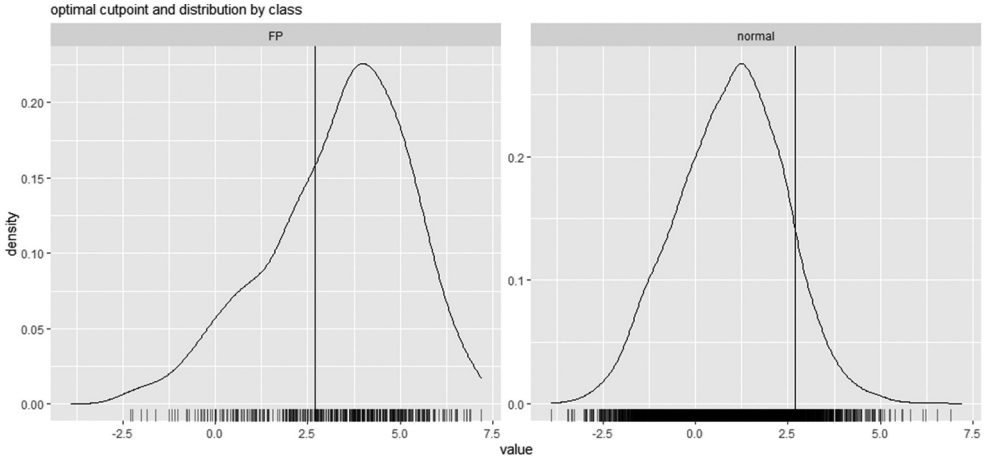

Figure 3. Density plots of scores resulting from the algorithm in the Dutch primary corpus training set, indicating optimal cutpoint with vertical line.

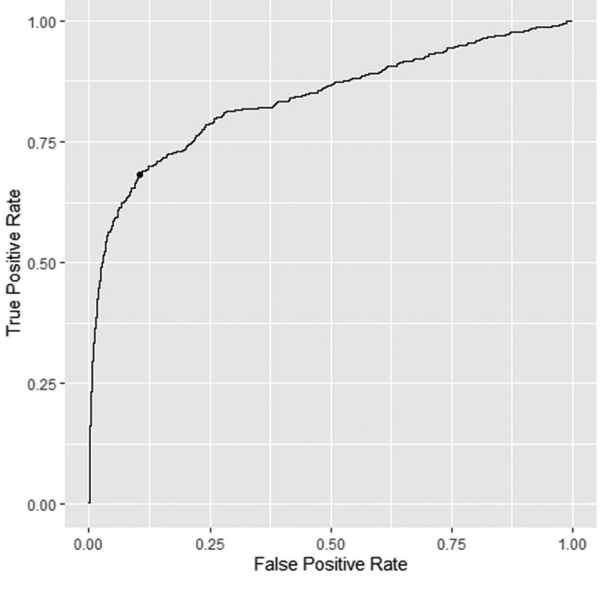

Figure 4. ROC curve of the scores in the Dutch primary corpus training set, indicating optimal point with lowest false positive rate and highest true positive rate (at threshold = 2.7094).

PRAAT script to detect filled pauses can be found in Appendix 3 (online). Figure 5 shows the TextGrid that results when running the rewritten PRAAT-script to detect syllable nuclei and the (embedded) new script to detect filled pauses.

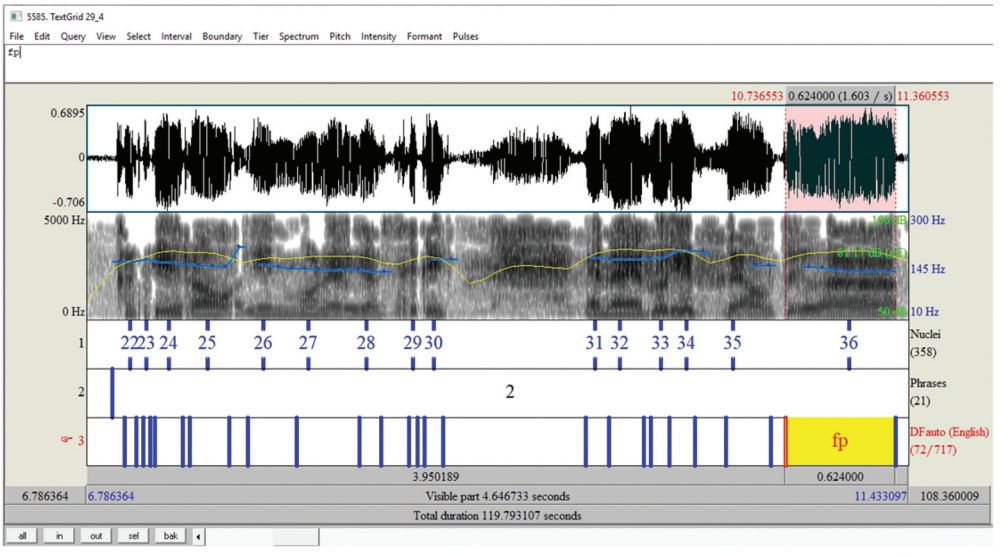

Figure 5. PRAAT screenshot of visualisation of a sound file with automatically-generated TextGrid; as can be seen in the third tier called 'DFauto', a filled pause ('fp') has automatically been detected, starting at 10.737 and ending at 11.361 seconds. The first tier shows the automatically detected syllable nuclei (22 through 36 in this soundfile), and the second tier shows the phrases as intervals (second phrase in this sound file).

Results

Testing accuracy of the algorithms

To test the accuracy of the algorithms for both corpora, using the function *confusionMatrix* from the package *caret* (Kuhn, 2019) in R, the relevant measures to compare how well the automatic function could detect the manually annotated filled pauses, were calculated in both test sets. These measures are calculated based on the filled-in confusion matrix. As an example, Table 3 shows this confusion matrix for the Dutch test data. There are a number of ways to report performance or accuracy of an algorithm on the basis of such a confusion matrix. The following measures are reported in Table 4:

- Accuracy = (true positives + true negatives)/total number = (107 + 1596)/1984;
- True positive rate or sensitivity = true positives/(true positives + false negatives) = 107/165;
- True negative rate or specificity = true negatives/(true negatives + false positives) = 1596/1819;

Table 3. Confusion matrix for Dutch test data (primary corpus, n = 1984).

		Manual annotation		
Automatic detection		FP	'normal'	totals
	FP	107	223	330
	'normal'	58	1596	1654
	totals	165	1819	1984

Table 4. Accuracy measures calculated on test data (30% of total) of primary corpora.

	Dutch test data (n = 1984)	English test data (n = 5935)
Accuracy	0.86	0.83
Sensitivity	0.65	0.56
Specificity	0.88	0.86
Precision	0.32	0.33
AUC	0.80	0.75

Table 5. Accuracy measures calculated on secondary corpora.

	Dutch L1 data (n = 24,006)	English L2 data (n = 22,366)	English L1 data (n = 4179)
Accuracy	0.90	0.84	0.86
Sensitivity	0.75	0.76	0.68
Specificity	0.91	0.85	0.87
Precision	0.28	0.21	0.20
AUC	0.88	0.87	0.87

- Precision = true positives/(true positives + false positives) = 107/330.

Finally, the measure area under the curve (AUC) is calculated from the earlier mentioned ROC-curve, but now using the test-data. This measure represents the extent to which the true positive rate (or sensitivity) is maximised while at the same time the false positive rate is minimised across all potential thresholds. This measure of performance is specifically useful for situations in which a logarithm is tested in skewed or unbalanced data sets (Fawcett, 2006), such as ours: most syllables are not filled pauses, only a few are.

Although the Dutch corpus was much smaller than the English corpus (which could have led to an unstable formula), most measures are slightly higher in the Dutch corpus compared to the English one. For instance, the measure 'sensitivity' (also called true positive rate or recall) is .65 for the Dutch test set, and .56 for the English data set. These numbers mean that 65% and 56% of all manually annotated filled pauses were also indicated as filled pauses by the automatic algorithm in the Dutch and English test sets, respectively. Most likely the lower scores for English are due to the fact that the quality of the recordings of the English corpus were worse than those of the Dutch corpus. Note, however, that the precision score for English is somewhat higher (ratio of true positives over all positives) than for Dutch (and both are quite low).

The results of testing accuracy of these algorithms on the secondary corpora can be found in Table 5. Interestingly, for the new Dutch L1 data, all accuracy measures are higher ('better') than those for the Dutch primary corpus on which the algorithm was trained, especially sensitivity which is up to 75%. For the new English data, we see that both L2 and L1 filled pauses can be predicted about equally well. Again, especially the scores for sensitivity (76% and 68%) are higher than the scores obtained from the test data of the primary corpus. However, note that the scores for precision are lower than precision scores for the primary L2 English corpus.

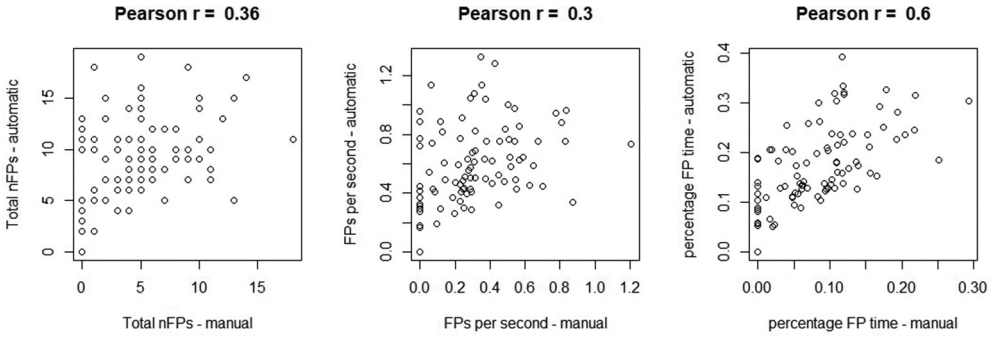

Figure 6. Scatter plots and Pearson correlations for the Dutch primary corpus, comparing measures from manually annotated and from automatically detected filled pauses (n = 90).

Correlations between manual and automatic measures of fluency

Global accuracy for the primary corpora was also gauged by correlating global measures of filled pauses as calculated from the manual annotations with the same global measures as calculated from the automatically detected filled pauses. For these analyses, the total number of filled pauses, their total durations, and total speaking time for each recording were measured. For the Dutch recordings from native speakers, these (manually measured) data were not available, thus leaving 90 Dutch recordings and 60 English recordings. Figure 6 (Dutch corpus) and 7 (English corpus) show the correlations and plots for the following measures, as calculated from the manual and automatic annotations: total number of filled pauses, number of filled pauses per second of speaking time, and percentage duration of all filled pauses of speaking time.

Figures 6 and 7 show that the correlations between automatic measures of filled pauses and manual measures of filled pauses are higher for the English data than for the Dutch data. A potential explanation for this finding is that the Dutch recordings (always around 20 seconds) were considerably shorter than the English recordings (mean 108 seconds, with a minimum of 37 seconds). It is quite likely that measures from recordings of just 20 seconds are less stable than those from longer recordings. We therefore also calculated the correlations adopting a stricter threshold for filled pauses in both data sets,

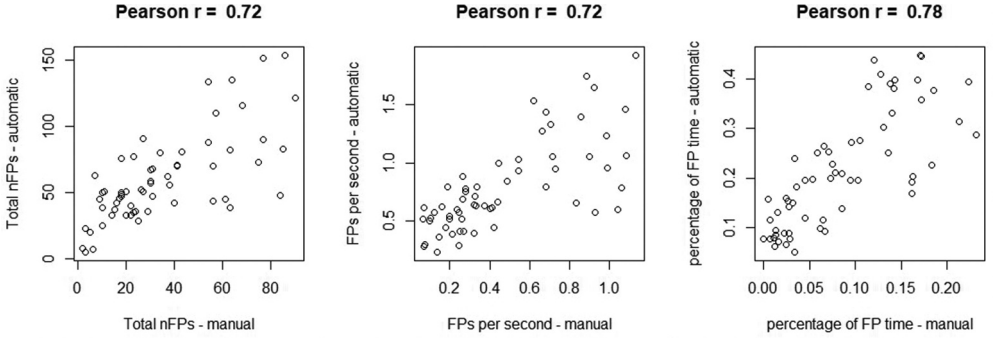

Figure 7. Scatter plots and Pearson correlations for the English primary corpus, comparing measures from manually annotated filled pauses with those from automatically detected filled pauses (n = 60).

Table 6. Total R^2 for linear models predicting (mean) scores for fluency.

Total R^2 Predictors	Dutch primary corpus (n = 90)	English primary corpus (n = 60)
Manual, only FP-measures	0.15	0.01 (ns)
Manual, including speech rate	0.75	0.43
Automatic, only FP-measures	0.16	0.02 (ns)
Automatic, including speech rate	0.53	0.32

multiplying the mathematically optimal cut points (found to be 2.7094 and 3.4942) by 1.2. Recalculating the relevant measure for filled pauses in this way led to higher correlations for both data sets, especially for the Dutch data set: correlations were now .57, .53. and .69 for the Dutch data and .78, .75, and .77 for the English data for the measures total number of filled pauses, number of filled pauses per second of speaking time, and percentage duration of all filled pauses of speaking time, respectively.

Gauging validity for the purpose of language assessment

In addition to validating the script in terms of accuracy of measurement compared to manual measurement, it is useful to evaluate the extent to which the automatic measures are predictive of human judgements on (aspects of) fluency and proficiency. As described in the materials sections, the speech data in both primary corpora have been judged on fluency.

For both data sets, we ran linear models to predict the fluency ratings by the scores for speech rate, number of filled pauses per second, and percentage duration of filled pauses as predictors. Table 6 shows the total R^2 for the models using the automatic measures and those for the models using the manual measures as predictors. We also gauged the predictive value of the new measures of filled pauses alone.

As can be seen from Table 6, the scores given by judges rating on fluency are best predicted when the measures for speech rate are included. For the English corpus, the measures of filled pauses (either by hand or automatically) do not predict these scores at all (the models were not significant). For the Dutch corpus, on the other hand, the models predicting the fluency scores by measurement of filled pauses alone, led to 15% (manually measured filled pauses) and 16% (automatically measured filled pauses) of the variance explained. This amount of variance explained by the automatic measures of filled pauses alone goes up to 20% when the higher cut point (multiplied by 1.2) was employed.

Discussion and conclusions

An algorithm to detect filled pauses automatically was created for Dutch and English speech data. The speech data of the primary corpora, on which the algorithms were trained, were mostly L2 data (Dutch corpus) or only L2 data (English corpus). Annotations by two annotators (Dutch data) or three annotators (English corpus) were combined in order to find the optimal algorithm to detect filled pauses. The algorithms for both corpora used almost the same set of variables, but with different coefficients. In general, it turned out that syllables

- that are long in duration,
- pronounced with a relatively low pitch,
- with vowel qualities that are like a schwa and/or like a back mid-open vowel,
- with vowel qualities that are stable,
- and pronounced lazily,

have a higher chance of being a filled pause. The algorithms and cut scores were determined using training data (70% of all syllables). Most of the measures included in the algorithms have been incorporated in previous research characterising and/or detecting filled pauses (e.g. Audhkhasi et al., 2009; Hughes et al., 2016b; Kaushik et al., 2010; Shriberg, 2001; Stouten & Martens, 2003; Vasilescu & Adda-Decker, 2007; Verkhodanova & Shapranov, 2016). For both the Dutch as well as the English algorithm, however, we found an additional characteristic to be predictive of syllables being a filled pause. This is summarised as 'pronounced lazily' above and pertains to the extent to which the speaker makes an effort in pronouncing the vowel, measured by the proximity of the syllable's F1 and F2 to the median of the speaker's F1 and F2. In short: the less effort a speaker makes, the closer the F1 and F2 are to the respective medians, and thus the more likely the syllable is a filled pause.

For the second research aim, we tested the accuracy of the algorithms on the remaining 30% of the data and found that 65% of the Dutch manually annotated filled pauses and 56% of the English manually annotated filled pauses were indeed correctly classified. At the same time, 12% (Dutch data) and 14% (English data) of syllables that were manually not annotated as being a filled pause, were incorrectly classified as being filled pauses (indicated by the specificity measures). Finally, precision for the current system seems quite low: 32% (Dutch data) or 33% (English data) of the automatically classified filled pauses are indeed as such annotated manually. Perhaps the high numbers of false positives can partly be explained by lengthened syllables that are now classified as filled pauses, a hypothesis we aim to investigate in future research. Related to this hypothesis is a question for future research, namely to what extent such false positives or lengthened syllables that partly sound like filled pauses could be classified as *hesitations*. For the second research aim, we also tested the algorithms on unrelated secondary corpora with Dutch L1 data, as well as English L2 and L1 data. The results showed that the accuracy measures for L1 Dutch were even higher than those of the primary corpus. For English, most accuracy measures were better, except the number of false positives ('precision' was lower than on the primary corpus).

In testing for the optimal algorithm, we decided to delete significantly predicting variables, if their sign in the formula was inconsistent with previous studies and theory. For instance, the standard deviation of F0 was found to add to predicting filled pauses. However, contrary to theory, it was found that *larger* standard deviations were related to higher probabilities of filled pauses, although previous research had indicated that stability of the F0 would be a positive predictor of filled pause occurrence (Verkhodanova & Shapranov, 2016). We concluded that such unpredicted results may have been due to correlation among the predictor variables. For F0, however, another possible explanation is that this was due to the fact that F0 may also tend to decline during the articulation of the filled pause (O'Shaughnessy, 1992). In other words, previous research is inconsistent with respect to F0 as a predictor of filled pauses:

a decline of F0 (O'Shaughnessy, 1992) would be consistent with larger standard deviations, whereas stability of F0 (Verkhodanova & Shapranov, 2016) would be consistent with smaller standard deviations. Therefore, decline of F0 should also be investigated as a potential indicator in future research.

In validating the automatic filled pause script further, global correlations between the manual and automatic measures of filled pauses, averaging over the recordings in the primary corpora, were carried out. The crucial global automatic measures, namely number of filled pauses per second speaking time and the percentage duration of filled pauses in speaking time, correlated with the manual measures between .53 and .78, at least when a more strict cut point or threshold was employed to detect filled pauses (the mathematically optimal found threshold multiplied by 1.2). These correlations were lower for the Dutch data than for the English data.

For the third research aim, we tested to what extent the algorithm can be used for L2-fluency measurement for assessment purposes, by comparing how well the manual and automatically measured aspects of fluency could predict human ratings of fluency. For these general linear models, only the Dutch measures of filled pauses significantly predicted these perception data, and the measures calculated from automatically detected syllables were performing at least as well as the measures calculated from manually detected syllables. It is possible that the judges rating the Dutch data had more precise instructions on how to judge fluency (including filled pauses) than the judges for the English data (compare Bosker et al., 2013 with Tavakoli et al., 2017). The difference in type of instruction may partly explain why the English ratings were not related to either automatic or manual measures of filled pauses. Another issue to keep in mind is that the current automatic measures of (filled) pauses do not take into account the location of pauses, while the study by Kahng (2018) has shown that location does play a role, at least for silent pauses: pauses within clauses are penalised more heavily than pauses between clauses. Therefore, if an automatic measurement of silent and filled pauses would be used in automatic scoring for language assessment, location should contribute in some way.

It is at this point difficult to evaluate whether the current algorithms are 'good' predictors for Dutch and English L2 filled pauses. First of all, there is variability in the way in which accuracy is reported and calculated across studies. For instance, Stouten and Martens (2003) report sensitivity of 70% for Flemish data and in the current Dutch data, 64% sensitivity was reported (see Table 5). However, note that the sensitivity for the current Dutch data test set would also increase to 70% if only those filled pauses that were marked by both annotators were to be counted. Secondly, more data would need to be tested: the English recordings of the primary corpus were of low quality for the purpose of acoustic analyses, and for the Dutch data in the primary corpus, only short recordings were included. However, the results from the accuracy measures of the secondary corpora seem promising, especially for the Dutch algorithm. As a next step, we propose to test the new PRAAT scripts on yet more data for both languages. With respect to the global validation (relating measures from the script to measures from manual annotations and to [fluency] ratings), we can be certain that these would be higher if the speaking excerpts (or the tasks themselves) had been longer. With respect to the sound quality, we can only speculate how this made a difference to the performance of the script. Possibly, the high number of false positives in the secondary corpus could be related to this. Another reason why it is hard to gauge the current

quality is that it is the first time an algorithm has been developed to detect filled pauses in L2 data. It is as yet unknown whether detecting filled pauses in L2 data may be more difficult than for L1 data. A hint is found in the positive results for the Dutch L1 data (secondary corpus), but this will be a challenge for future research, along with the question of whether correctly detecting filled pauses may be dependent on L2 proficiency. It may well be that as proficiency progresses, filled pauses sound more like those from native speakers, and hence will be more often (correctly) detected than from beginner learners. Taking these considerations into account, we must conclude that the current algorithm is not yet suitable for automatic measurement of fluency for the purpose of assessing learners' proficiency, for instance, for the APTIS speaking test. Future research should test not only whether overall proficiency would play a role in the accuracy of (and therefore validity to use) the script, but also whether factors such as gender, quality and length of sound recording, L1 background of the speaker may play a role and should be systematically tested. Additionally, it is worth researching whether judges who rate on only one aspect (e.g. fluency as in the Dutch corpus) are able to focus more precisely on this factor alone compared to judges who make a holistic assessment alongside analytic ratings on specific aspects (as was the case in the English corpus).

To conclude, the first aim of this research was to create an easy-to-use PRAAT script that can measure aspects of L2 fluency automatically, without the time-consuming task of transcribing the speech beforehand. For this purpose, the De Jong and Wempe (2009) script has been updated and rewritten, and syllables are now detected with more accuracy; this is evidenced by a higher correlation between articulation rate calculated from manually counted syllables with articulation rate calculated by the script. A new script was subsequently created to automatically detect filled pauses. With respect to the second aim, to test the accuracy of the new script, we can conclude that the current algorithm can correctly identify most of the manually annotated filled pauses, even in corpora on which the algorithms were not trained. At the same time, however, there is currently a high number of false positives, i.e. syllables that the algorithm identifies as filled pauses, but which were not manually classified as such. Finally, regarding the third aim, if judges are instructed to look out for fluency, with a precise definition that includes filled pauses as was the case for the Dutch data, up to 20% of the variance of ratings on fluency can be explained by the automatically measured filled pauses alone. In conclusion, although further research in validation is necessary, the current script with the algorithm is promising, especially given the stable performance on the secondary corpora.

Notes

1. The speaking tasks 1 through 3 in APTIS were all scored between 0 through 5, where a score of 5 entails that the performance is above the level of the task; For task 4, score 5 is at C1-level, and 6 is defined as above C1 level. Participants at the C (at least C1) level in Tavakoli et al. (2017) always scored at least a 5, on all tasks.
2. The pitch ceiling in this step is based on the estimates of mean 110 and 190 Hz for male and female speech and the usual pitch ceilings of 225 and 450 yield a factor between 2 and 2.4. Our choice of 2.5 is therefore chosen as a safe multiplier to determine pitch ceiling. (Rietveld & Van Heuven, 2009).

3. The PRAAT manual advises 5500 Hz for measuring 5 formants in female speech and usually 100 Hz is seen as a good choice per formant. Rietveld and Van Heuven (2009) note that for male speech, formant values are about 10% lower. The current formula approximates the thus obtained max values of 4400 and 4000 Hz for female and male speech when measuring four formants (which is good practice when three are required).

Acknowledgments

We are grateful to The British Council Assessment Research Awards and Grants programme 2018 for funding this project, providing part of the data, as well as feedback on previous versions of this paper. In addition, the third author was supported by the Netherlands Organization for Scientific Research (NWO VIDI grant 276-75-010).

Research ethics

We used three existing speech corpora. For each corpus, informed consent was received from all participants. All corpus data was anonymised and used in a secure and confidential manner in line with principles of research ethics codes.

Disclosure statement

No potential conflict of interest was reported by the author(s).

Funding

This work was supported by The British Council Assessment Research Grants (first and second authors) and by the Netherlands Organisation for Scientific Research [276-75-010] (third author).

Notes on contributors

Nivja H. de Jong is Associate Professor Second Language Acquisition and Pedagogy at LUCL (Leiden University Center of Linguistics) and at ICLON (Leiden University Graduate School of Teaching). She is chair of the LLRC (Language Learning Resource Centre) at Leiden University and is an expert member of EALTA.

Jos Pacilly is Senior Electronics Engineer and Software Developer at the Phonetics Laboratory of LUCL (Leiden University Center of Linguistics).

Willemijn Heeren is Associate Professor Forensic Phonetics at LUCL (Leiden University Center of Linguistics) and Principal Investigator of 'The Speaker in Speech - the interdependence of linguistics and indexical information' (2017-2022, NWO VIDI 276-75-010).

ORCID

Nivja H. de Jong http://orcid.org/0000-0002-3680-3820

References

Audhkhasi, K., Kandhway, K., Deshmukh, O. D., & Verma, A. (2009). Formant-based technique for automatic filled pause detection in spontaneous spoken English, *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, (pp. 4857–4860). IEEE.

Belsley, D. (1991). *Conditioning Diagnostics: Collinearity and Weak Data in Regression*. Wiley.

Boersma, P., & Weenink, D. (2016). PRAAT: Doing phonetics by computer [Computer program]. Version 6.1.29. Retrieved 28 October 2019, from http://www.PRAAT.org/ .

Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, *30*(2), 159–175. https://doi.org/10.1177%2F0265532212455394

Clark, H. H., & Fox Tree, J. E. (2002). Using uh and um in spontaneous speaking. *Cognition*, *84*(1), 73–111. https://doi.org/10.1016/s0010-0277(02)00017-3

De Boer, M. M., & Heeren, W. F. L. (2020). Cross-linguistic filled pause realization: The acoustics of uh and um in native Dutch and non-native English. *Journal of the Acoustical Society of America*, *148*(6), 3612–3622. https://doi.org/10.1121/10.0002871

De Jong, N. H. (2016). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, *54*(2), 113-132. https://doi.org/10.1515/iral-2016-9993

De Jong, N. H., & Wempe, T. (2009). PRAAT script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods*, *41*(2), 385–390. https://doi.org/10.3758/BRM.41.2.385

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *27*(8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics*, *21*(3), 354–375. https://doi.org/10.1093/applin/21.3.354

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral english proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379–399. https://doi.org/10.1177/0265532210364407

Hughes, V., Foulkes, P., & Wood, S. (2016a). Formant dynamics and durations of um improve the performance of automatic speaker recognition systems. *Proceedings of the 16th Australasian Conference on Speech Science and Technology (ASSTA)*. Western Sydney University

Hughes, V., Wood, S., & Foulkes, P. (2016b). Strength of forensic voice comparison evidence from the acoustics of filled pauses. *International Journal of Speech, Language and the Law*, *23*(1), 99–132. https://doi.org/10.1558/ijsll.v23i1.29874

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*(2), 202–220. https://doi.org/10.1177/0265532211419826

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied Linguistics*, *29*(1), 24–49. https://doi.org/10.1093/applin/amm017

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning with applications in R*. Springer.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 english speakers: Temporal measures and stimulated recall. *Language Learning*, *64*(4), 809–854. https://doi.org/10.1111/lang.12084

Kahng, J. (2018). The effect of pause location on perceived fluency. *Applied Psycholinguistics*, *39*(3), 569–591. https://doi.org/10.1017/S0142716417000534

Kang, O., Rubin, D., & Pickering, L. (2010). Suprasegmental measures of accentedness and judgments of language learner proficiency in oral english. *The Modern Language Journal*, *94*(4), 554–566. https://doi.org/10.1111/j.1540-4781.2010.01091.x

Kaushik, M., Trinkle, M., & Hashemi-Sakhtsari, A. (2010). Automatic detection and removal of disfluencies from spontaneous speech. *Proceedings of the 13th Australasian International Conference on Speech Science and Technology*, (pp. 98–101).

Krikke, T. F., & Truong, K. P. (2013). Detection of nonverbal vocalizations using gaussian mixture models: Looking for fillers and laughter in conversational speech. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH* (pp. 163–167). International Speech and Communication Association.

Kuhn, M. (2019). caret: classification and regression training. R package version 6.0-84. https://CRAN.R-project.org/package=caret .

Ladefoged, P., & Johnson, K. (2011). *A course in phonetics*. Wadsworth.

Levelt, W. J. M., Roelofs, A., & Meyer, A. S. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*(1), 1–37. https://doi.org/10.1017/s0140525x99001776

Moussalli, S., & Cardoso, W. (2020). Intelligent personal assistants: Can they understand and be understood by accented L2 learners? *Computer Assisted Language Learning*, *33*(8), 865–890. https://doi.org/10.1080/09588221.2019.1595664

O'Shaughnessy, D. (1992). Recognition of hesitations in spontaneous speech. *Proceedings ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing (521-524)*. IEEE. https://doi.org/10.1109/ICASSP.1992.225857

Orr, R., & Quené, H. (2017). D-LUCEA: Curation of the UCU accent project data. In J. Odijk & A. van Hessen (Eds.), *CLARIN in the low countries* (pp. 177–190). Ubiquity Press.

Reetz, H., & Jongman, A. (2009). *Phonetics: Transcription, production, acoustics, and perception*. Wiley-Blackwell.

Révész, A., Ekiert, M., & Torgersen, E. N. (2016). The effects of complexity, accuracy, and fluency on communicative adequacy in oral task performance. *Applied Linguistics*, *37*(6), 828–848. https://doi.org/10.1093/applin/amu069

Rietveld, T., & Van Heuven, V. J. (2009). *Algemene Fonetiek (third edition) [General phonetics]*. Coutinho.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Shriberg, E. (2001). To 'errrr' is human: Ecology and acoustics of speech disfluencies. *Journal of the International Phonetic Association*, *31*(1), 153–169. https://doi.org/10.1017/S0025100301001128

Shriberg, E. E., & Lickley, R. J. (1993). Intonation of clause-internal filled pauses. *Phonetica*, *50*(3), 172–179. https://doi.org/10.1159/000261937

Stouten, F., Duchateau, J., Martens, J. P., & Wambacq, P. (2006). Coping with disfluencies in spontaneous speech recognition: Acoustic detection and linguistic context manipulation. *Speech Communication*, *48*(11), 1590–1606. https://doi.org/10.1016/j.specom.2006.04.004

Stouten, F., & Martens, J. P. (2003). A feature-based filled pause detection system for Dutch. In *2003 IEEE Workshop on Automatic Speech Recognition and Understanding*, (pp. 309–314. IEEE.

Tavakoli, P., Nakatsuhara, F., & Hunter, A. M. (2017). Scoring validity of the aptis speaking test: Investigating fluency across tasks and levels of proficiency. *ARAGs Research Reports Online*. 2057-5203. https://www.britishcouncil.org/sites/default/files/tavakoli_et_al_layout.pdf

Thiele, C., Wing, J., Weston, S., Williams, A., Keefer, C., Engelhardt, A., Cooper, C., Mayer, Z., Kenkel, B., Benesty, M., Lescarbeau, R., Ziem, A., Scrucca, L., Tang, Y., Candan, C., & Hunt, T. (2019). cutpointr: Determine and evaluate optimal cutpoints in binary classification tasks. R package version 0.7.6. https://CRAN.R-project.org/package=cutpointr .

Vasilescu, I., & Adda-Decker, M. (2007). A cross-language study of acoustic and prosodic characteristics of vocalic hesitations. In A. Esposito, M. Bratanić, E. Keller, & M. Marinaro (Eds.), *Fundamentals of verbal and nonverbal communication and the biometric issue* (pp. 140–148). IOS Press.

Verkhodanova, V., & Shapranov, V. (2016). Experiments on detection of voiced hesitations in Russian spontaneous speech. *Journal of Electrical and Computer Engineering*, *2016*(8), Article ID 2013658. https://doi.org/10.1155/2016/2013658