Choosing a threshold for silent pauses to measure second language fluency

Nivja H. De Jong & Hans Rutger Bosker

Utrecht Institute of Linguistics OTS, The Netherlands

Abstract

Second language (L2) research often involves analyses of acoustic measures of fluency. The studies investigating fluency, however, have been difficult to compare because the measures of fluency that were used differed widely. One of the differences between studies concerns the lower cut-off point for silent pauses, which has been set anywhere between 100 ms and 1000 ms. The goal of this paper is to find an optimal cut-off point. We calculate acoustic measures of fluency using different pause thresholds and then relate these measures to a measure of L2 proficiency and to ratings on fluency.

Index Terms: silent pauses, number of pauses, duration of pauses, silent pause threshold, second language speech.

1. Introduction

In research on both native (L1) and L2 speech, silent pauses are an important feature to describe, characterize, and compare speech from different speakers, performing in different speaker tasks. However, there is a longstanding debate on what should count as a pause. In connected spontaneous speech, part of the speech signal involves silence every time an occlusive is produced. These silences in speech are not considered as pauses that reflect hesitation behavior and it has been assumed that the silences before occlusives can quite easily be removed from a silent pause count by setting a certain threshold. Goldman-Eisler (1968) proposes a threshold of 250 ms to distinguish between 'articulatory' (<250 ms) and 'hesitation' (>250 ms) pauses [1], and this threshold has been followed both in research on L1 and L2 speech.

More recently, however, using this boundary has been called into question [2, 3, 4]. Most of pauses within the 130 ms – 250 ms range cannot be attributed to articulation [2]. Pauses as short as 60 ms that are *not* part of occlusives have been reported [3].

In L2 research, applying a threshold to measure number and duration of pauses has also been used. Often, Goldman-Eisler is cited and the boundary of 250 ms is used [5, 6]. But in some studies, a lower cut-off point is used [7: 100 ms], or a higher cut-off point is used [8: 400 ms], even as high as 1000 ms [9].

The current paper is an attempt to find the optimal cut-off point for the purpose of L2 research. We use two different strategies to find an optimal cut-off point. In L2 research, acoustic measures of fluency (such as number of silent pauses or speech rate) are used to compare speakers or performances of the same speakers in different tasks. These measures are thought to reflect automaticity in the L2 speech production processes. To find out which measures of fluency are indeed related to automaticity of L2 speech production, and to overall L2 proficiency, studies have related acoustic measures of fluency to subjective ratings on L2 proficiency [10], to subjective ratings on fluency [5, 11], but also to separate measures of L2 proficiency [6, 12].

In this paper, we calculate acoustic measures of fluency using different pause thresholds as lower cut-off points. We then evaluate (1) the relation between these measures of fluency and a measure of vocabulary knowledge as an approximate of overall L2 proficiency, and (2) the relation between the acoustic measures of fluency and subjective ratings. If we find that choosing a specific threshold leads to higher correlations either with L2 proficiency and/or with subjective fluency, this would argue for using this specific threshold in future L2 research on fluency.

Kirsner, Dunn and Hird [4] show that each individual may have his own criterion when distinguishing between short and long pauses, even fluctuating according to variables such as topic, task, time of day, and age. In the current paper, we will therefore also test whether a threshold that fluctuates per individual or speech sample improves the correlations between acoustic measures of fluency on the one hand, with L2 proficiency and perceived fluency, on the other.

2. Method

In what follows, we describe the data that were used in the present study. In short, the full corpus as described in [12] was used to evaluate the effect of different silent pause thresholds on the relation between acoustic measures of fluency with L2 proficiency (vocabulary knowledge); a subset of this corpus was used to evaluate the effect of different silent pause thresholds on the relation between acoustic measures of fluency with ratings.

2.1. Speech data

The corpus consisted of all L2-data from [12]. Fifty-one L2 speakers (24 Turkish L1 and 27 English L1) of Dutch performed eight speaking tasks in their L2. The total duration of speech in this L2-corpus was 9 hours and 43 minutes. For this data, orthographic transcriptions were made in CLAN [13]. Furthermore, silent pauses were detected by careful listening and by using the waveform (as shown in CLAN), and measured in milliseconds. The silent pauses were also classified with respect to their location, specifically whether the silent pauses occurred either within or between Analysis of Speech (AS) units [14]. AS-units can be described as utterances consisting of an independent clause or of a subclausal unit, together with the associated subordinate clause(s). In this study, we will report on measures of fluency based on pauses within AS units only.[1] In total, 10668 silent pauses within AS-units were identified.

Figure 1 shows the distribution of pause durations, after (natural) log transformation. Both [3] and [4] report most pauses to be falling in the "short pause" distribution (roughly under 200 ms), whereas in our distribution most pauses are longer.

[1] Calculating the acoustic measures using all silent pauses, rather than only those within AS-units, led to lower correlations across all analyses.

Our participants speak in their L2, which has probably caused a different distribution as found before in read and spontaneous L1 speech (as reported before in [3,4]). Secondly, pauses in our data were detected manually (the noise in our speech files was too variable to allow for automatic pause detection). Manual detection will lead to fewer very short pauses as compared to automatic detection.

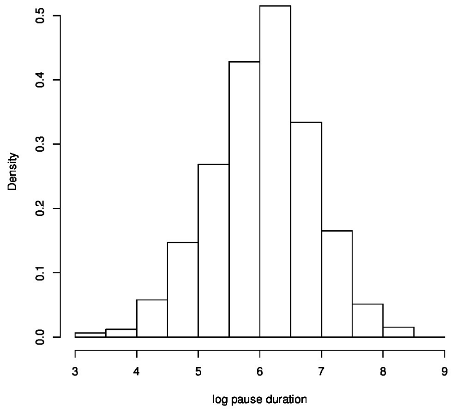

Figure 1: *Histogram of all pause durations in the (full) corpus.*

2.2. Perception data

A subset of the speech corpus was created for a listening experiment, as described in [11]. From 30 L2 speakers (15 L1 Turkish and 15 L1 English), speaking performances from three tasks were used. Twenty-second excerpts were taken from roughly the middle of these 90 speaking performances.

Twenty raters judged the speech samples on fluency on a 9-point equal appearing interval scale. From these ratings, so-called estimates were calculated. These estimates may be considered as mean ratings from the twenty judges, taking into account individual differences of raters, a general effect of order of presentation (raters became more strict towards the end of the experiment), and individual differences between raters with respect to this general order effect. For a complete description of how the perception data were obtained, we refer to [11]. The mean, sd, and range of these rating measures were 5.33, 1.51, and 1.34–8.50, respectively.

2.3. L2 proficiency data

In addition to performing the speaking tasks, the participants also carried out a productive vocabulary task with 116 items. We will use the scores on this task as a separate measure of L2 proficiency. Vocabulary knowledge has been shown to be a good predictor of overall proficiency [15]. Moreover, the same vocabulary test as was used in the current paper, has been shown to be a strong predictor of overall speaking proficiency [16]. The mean, sd, and range of these vocabulary scores were 56, 23, and 8 – 103, respectively.

2.4. Calculating fluency measures

We thus obtained two speech datasets: the full corpus of 51 speakers performing 8 speaking tasks (almost 10 hours of speech), and the subset of 90 roughly 20-second excerpts from thirty of these speakers (54 minutes of speech). To test what the impact may be of setting different thresholds of silent pauses on conclusions drawn in L2 fluency research, we calculated three acoustic measures of fluency that are strongly influenced by choosing different silent pause thresholds.

Choosing different thresholds will strongly influence some acoustic measures of fluency, but other measures of fluency will only change slightly, and for yet other measures of fluency, changing the threshold will not lead to any changes. For instance, speech rate, which is calculated by dividing number of syllables or number of phonemes by total time (including silent pauses), will not change depending on the chosen silent pause threshold because neither the number of syllables nor the total time will change. However, all measures of fluency that are calculated relative to phonation time will slightly change depending on the silent pause threshold, because the phonation time will change accordingly.

In the current paper, we focus on measures of fluency for which changing the silent pause threshold may lead to larger differences: i.e., we focus on the number and duration of silent pauses. For each participant (N = 51) or for each speech sample (N = 90), we calculated three measures of fluency: number of silent pauses per second total time, number of silent pauses per second phonation time, and mean (log) duration of silent pauses. These calculations were made using thresholds for the lower boundary of silent pause durations: at 20, 50, 100, and then at every 50 ms up to 1000 ms for the full corpus. The calculations for the smaller corpus of 90 20-second speech segments were made using the same thresholds, but in these samples the highest threshold was set to 400 ms, because higher thresholds would lead to missing data points (rendering comparisons across thresholds impossible).

Table 1 shows correlations between the fluency measures at a quite low (50 ms) and quite high (400 ms) threshold. It is not surprising that the correlation between the two frequency measures number of pauses per total time and number of pauses per phonation time are highly related (at both thresholds $r = 0.95$). At a boundary of 400 ms, we also see a strong correlation between number of pauses total time and mean duration of pauses ($r = 0.52$).

The correlations across the two thresholds were also carried out (not shown in Table 1). For the measure mean log duration, the measure calculated at 50 ms and 400 ms is highly related ($r = 0.95$), whereas for the two frequency measures the correlation between the measures calculated at the two different thresholds is less strong ($r = 0.65$ and $r = 0.60$ for the frequency measures calculated per total time and per phonation time, respectively). We can therefore expect that for the analyses in which we relate the measures of fluency to vocabulary knowledge and to ratings on fluency, we will not find changes for mean pause duration, as this measure hardly changes with different thresholds. For the frequency measures, on the other hand, we may expect to find differences.

Table 1: *Correlations between fluency measures in the ful corpus for measures calculated at thresholds 50 m and 400 ms.*

	50 ms			400 ms		
	1	2	3	1	2	3
Pauses / sec total time (1)	1	.95	−.06	1	.95	.52
Pauses / sec phon time (2)		1	−.33		1	.30
Log duration of pauses (3)			1			1

As stated in the Introduction, we also calculated measures of fluency with individualized thresholds. Kirsner and colleagues report individualized thresholds with a mean, standard deviation and range of 255, 83 and 98–490 ms, respectively [4]. They established these individualized thresholds by modeling the individual distributions with bi-Gaussian fits per individual. We will calculate individual thresholds in a different way, for two reasons. The first reason is that our data do not follow clear bi-Gaussian distributions The second reason is that in our data we have information on articulation rates available: the faster the articulation rate, the shorter the articulation pauses must be. To calculate individual thresholds, we will therefore use a threshold for each individual (or speech sample) that is relative to the individual's articulation rate. For the full corpus, the mean threshold was set to 250 ms, and, relative to individual's articulation rate, an individualized threshold was calculated, with a range of individualized thresholds between 139–324 ms. For the small corpus, the individualized thresholds were also around 250 ms, ranging from 138–384 ms, now relative to each of the 90 speech samples' articulation rates.

3. Analyses

3.1. Relating measures of L2 fluency to L2 proficiency (thresholds 20 ms – 1000 ms)

We related the acoustic fluency measures, as calculated from the full corpus described above, to the measure of L2 proficiency (vocabulary knowledge). Figure 2 shows the Pearson correlations (on the y-axis) between the measure of L2 proficiency and the fluency measures (as shown by different lines), for the silent pause thresholds 20 ms – 1000 ms (on the x-axis).

For mean log pause duration, none of the Pearson correlations were found to be significant. For the two frequency measures of pauses, there is a rise in correlations from thresholds 20 ms (per total time: $r = –0.42$; per speaking time: $r = –0.39$) to 300 ms ($r = –0.48$ and $r = –0.53$, respectively), after which the correlations drop.

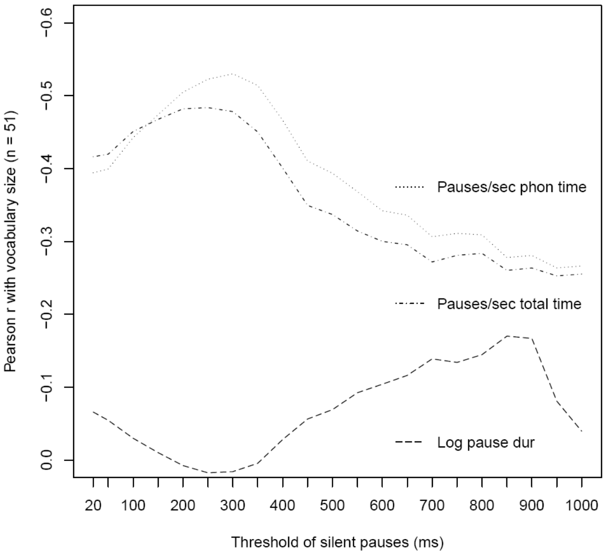

Figure 2: *Pearson correlations between vocabulary size and measures of fluency, calculated for different silent pause cut-off points.*

3.2. Relating measures of L2 fluency to perceived fluency (thresholds 20 ms – 400 ms)

For each speech sample of roughly 20 seconds ($N = 90$) in the small corpus, we calculated the acoustic measures of fluency as described above. Figure 3 shows the Pearson correlations (on the y-axis) between the measure of perceived fluency and the three acoustic fluency measures (as shown by different lines), for the silent pause thresholds 20 ms – 400 ms (on the x-axis).

As can be seen from this figure, changing the threshold from 20 ms to 400 ms does not lead to differences in correlations between the measure log pause duration and perceived fluency. For both frequency measures of pauses, however, we find that the higher the lower cut-off point for silent pauses, the higher the correlation between the resulting frequency measure (either number of silent pauses per total time or number of silent pauses per phonation time) on the one hand, and the ratings of fluency, on the other.

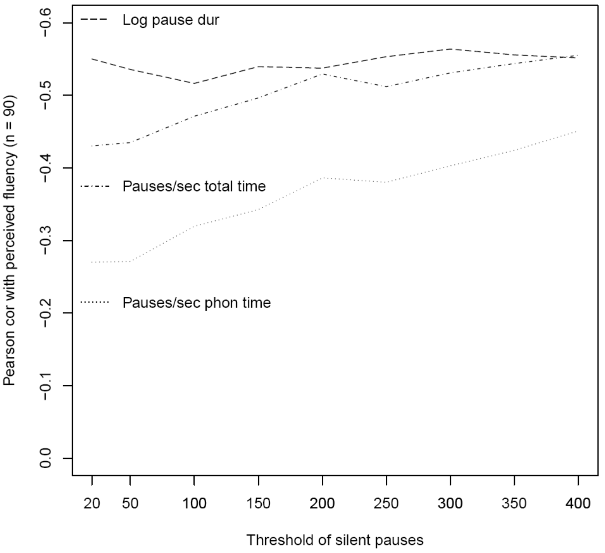

Figure 3: *Pearson correlations between ratings of fluency and measures of fluency, calculated for different silent pause cut-off points.*

3.3. Applying individual thresholds

As described above, we also calculated the measures of fluency when using individualized thresholds, relative to the articulation rates. For the correlations between acoustic measures of fluency and vocabulary knowledge, using these individualized thresholds, did not lead to different results from using the non-individual thresholds of 250 ms or 300 ms.

For the correlations between acoustic measures of fluency and the ratings, we did find differences, however: when number of pauses (per phonation time and per total time) was calculated using the individualized thresholds, the Pearson correlations were $r = –0.49$ and $r = –0.60$ respectively. These are higher than the correlations found for a non-individual threshold of 250 ms ($r = –0.38$ and $r = –0.51$ respectively). The correlation between perceived fluency and log duration of pauses when calculated with the individualized threshold, on the other hand, was lower ($r = –0.47$) than when it was calculated with a non-individual threshold of 250 ms ($r = –0.55$).

4. Discussion

From our results, a number of observations can be made. The first is that the correlations between log duration of pauses and vocabulary knowledge are always low and never significant (around $r = –0.1$; see Figure 2). The correlations between log duration of pauses and perceived fluency, on the other hand, are always much higher ($r = –0.55$; see Figure 3), irrespective of the threshold. This general discrepancy (low correlation between duration of pauses for measures of proficiency and high correlation for ratings of fluency) has been reported before [6,11]. What we can conclude from the present study, however, is that these findings are not dependent on a specific threshold.

Another finding from the current study is that the relation between vocabulary size and number of silent pauses *is* dependent on the chosen threshold. This relation is highest when a threshold of around 250–300 ms is used. We may conclude from this finding, that 250–300 ms is the optimal threshold for measuring the number of pauses (per total or per speaking time) with respect to studies that aim to investigate L2 proficiency. In other words, adding the number of pauses below 250 ms to counts obtained when a traditional cut-off point of 250 ms is used, leads to a measure of fluency that is less strongly related to L2 proficiency. Similarly, setting the threshold higher than 300 ms leads to lower correlations. We conclude from this finding that although many silent pauses are shorter than 250 ms (in our data between 22% and 27%), these pauses seem irrelevant when calculating measures of fluency that are related to L2 proficiency.

Such an optimal threshold for the number of silent pauses could not be found when relating the measures to perceived fluency; in this case the correlations get stronger as the threshold is higher. We could conclude from this finding that raters only take the number of long pauses into account (at least >400 ms) when judging on fluency. However, we propose another explanation for this finding. It follows naturally that a count of silent pauses with a high threshold is related to mean duration of silent pauses. If only the number of long pauses is counted, this count will be strongly related to the mean duration of pauses. Indeed, the correlation between mean log duration of silent pauses and the number of silent pauses when using a threshold of 400 ms is quite high (see Table 1: $r = 0.52$) and gets steadily higher if the threshold for counting pauses per second is raised (to $r = 0.77$ for a threshold 750 ms). The rise in correlations between number of pauses and ratings on fluency as the threshold is set higher, can therefore be explained by the fact that counting only long pauses is confounded with measuring mean duration of silent pauses (a measure that was strongly related to ratings on fluency).

In this study, we have also compared two frequency measures of pauses: number of pauses per second total time and number of pauses per second phonation time. For L2 proficiency, the correlations were almost the same for these measures (slightly higher when it was calculated per second phonation time). For the ratings, however, the correlations were higher when the acoustic measure was calculated per total time. We can explain this finding, again, by taking into account the intercorrelations between the acoustic fluency measures: the number of pauses per total time, especially as the threshold gets higher, is in fact a confounded measure of the number of pauses and their duration.

5. Conclusions

This study showed that a lower cut-off point for silent pauses of 250–300 ms leads to the highest correlation between the number of silent pauses and a measure of L2 proficiency (vocabulary knowledge). Such an optimal threshold could not be found for the mean (log) duration of silent pauses in relation to L2 proficiency: mean duration of silent pauses is not significantly related to L2 proficiency, no matter which threshold was chosen.

When relating the acoustic fluency measures to ratings on fluency, no clear optimum could be found. For mean duration of pauses, correlations between ratings and this measure were always high, irrespective of the threshold. For the number of silent pauses, the correlations became higher, as the threshold was set higher. This finding, however, can be explained by the fact that counting only long pauses (by setting the threshold high) is confounded with measuring the duration of pauses.

We therefore conclude that for the purpose of L2 research, the traditional cut-off point of 250 ms is a good choice and using a higher threshold than 300 ms has two disadvantages: (1) with respect to number of pauses, it leads to measures of fluency that are less representative of L2 proficiency, and (2) the acoustic fluency measures number of pauses and duration of pauses become confounded as higher thresholds are used.

6. Acknowledgements

The authors wish to thank research assistants Cem Keskin and Erica Bouma for measuring silent pauses; Cornelia Lahmann and Rasmus Steinkrauss for valuable discussion. This research was funded by Pearson Language Testing, grant "Oral fluency: production and perception" awarded to N.H. De Jong.

7. References

[1] F. Goldman-Eisler. *Psycholinguistics: Experiments in spontaneous speech*. New York: Academic Press, 1968.

[2] A.E. Hieke, S. Kowal and D.C. O'Connell. "The trouble with 'articulatory' pauses", *Language and Speech* 26, pp. 203–214, 1983.

[3] E. Campione and J. Véronis. "A large-scale multilingual study of silent pause duration." *Proc. ESCA-workshop*, pp. 199–202, 2002.

[4] K. Kirsner, J. Dunn and K. Hird. "Fluency: Time for a paradigm shift", *Proc. DiSS '03*, pp. 13–16, 2003.

[5] C. Cucchiarini, H. Strik and L. Boves. "Quantitative assessment of second language learners' fluency: Comparisons between read and spontaneous speech", *JASA* 111, pp. 2862–2873, 2010.

[6] N.H. De Jong, M.P. Steinel, A.F. Florijn, R. Schoonen and J.H. Hulstijn, "Linguistic skills and speaking fluency in a second language", *Applied Psycholinguistics*, online 2012.

[7] A. Riazantseva, "Second language proficiency and pausing: A study of Russian speakers of English", *SSLA* 23, pp. 497–526, 2001.

[8] T.M. Derwing, M.J. Munro, R.I. Thomson and M.J. Rossiter, "The relationship between L1 fluency and L2 fluency development", *SSLA* 31, pp. 533–557, 2009.

[9] N. Iwashita, "Features of oral proficiency in task performance by EFL and JFL learners". in M.T. Prior [Ed], Selected proceedings of the 2008 Second Language Research Forum*,* pp. 32–47, Cascadilla Proceedings Project, 2010.

[10] N. Iwashita, A. Brown, T. McNamara and S. O'Hagan, "Assessed Levels of Second Language Speaking Proficiency: How Distinct?", *Applied Linguistics* 29(1), pp. 24–49, 2008.

[11] H.R. Bosker, A.F. Pinget, H. Quené, T. Sanders and N.H. De Jong, "What makes speech sound fluent? The contributions of pauses, speed and repairs", *Language Testing* 30, pp. 159–175, 2013.

[12] N.H. De Jong, R. Groenhout, R. Schoonen and J.H. Hulstijn, "Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior." *Applied Psycholinguistics*, online 2013.

[13] B. MacWhinney, "The CHILDES project: Tools for analyzing talk". Mahwah, NJ: Erlbaum, 2000.

[14] P. Foster, A. Tonkyn and G. Wigglesworth, "Measuring spoken language: A unit for all reasons". *Applied Linguistics* 21, pp. 354–375, 2001.

[15] A. Zareva, P. Schwanenflugel and Y. Nikolova, "Relationship between lexical competence and language proficiency", *SSLA* 27, pp. 567–595, 2005.

[16] N.H. De Jong, M.P Steinel, A. Florijn, R. Schoonen and J.H. Hulstijn, "Facets of speaking proficiency", *SSLA* 34, pp. 5–34, 2012.