HUMAN PAUSE DETECTION IN SPONTANEOUS SPEECH IN AN ONLINE EXPERIMENT

Jürgen Trouvain, Raphael Werner

Language Science and Technology, Saarland University, Saarbrücken, Germany
trouvain|rwerner@lst.uni-saarland.de

Abstract: The aim of this exploratory study is to check whether listeners are able to detect the same pauses in spontaneous speech and how fast they are at doing so. The test with German data was performed online with students with basic skills in annotation. The results show that long pauses with breath noises were detected better than hesitation pauses. An important outcome is that pauses were also detected before silences occur or with stimuli without silences.

1 Introduction

Pausing in speech serves several important functions. Most prominently, pauses serve as markers of syntactic-prosodic phrase boundaries in spoken text. They also represent parts of hesitations in read and spontaneous speech. In addition, pauses are core elements of turn management in conversations.

The perception of pauses in speech depends on various factors for different levels of linguistic and phonetic processing (cf. [1, 2, 3, 4]): silence, pause-internal particles (e.g., inbreath noise or filler particles such as *uh(m)*), phrase-final lengthening, intonational boundary tones, voice quality (e.g., creaky voice), drops in intensity, syntactic information.

This contribution reports an online experiment aimed at the detection of pauses in spontaneous speech. It evaluates the strength of agreement between listeners detecting pauses in different conditions including pauses with and without inbreath noises, pauses with hesitations, and pauses without silences but with intonational and syntactic cues.

The main interest is formulated with three research questions:

1. Do listeners agree in detecting pauses at the same locations in stretches of spontaneous speech?

2. What cues do listeners rely on the most for the detection of speech pauses?

3. How fast are listeners in detecting those pauses?

The study has an exploratory character in two aspects: First, we check whether it is possible to gain usable data with the applied technique for such a detection test. Due to the pandemic situation it was not possible to ask subjects to perform a listening experiment in a lab with the presence of an experimenter controlling for comparable procedures. As a remote testing environment we decided to use Praat [5] with the annotation function of so-called points (instead of marking starts and ends of a given stretch). Such a procedure requires some basic knowledge of how to annotate in Praat – which is the case for students of phonetics and other speech sciences.

The second exploration concerns the length and the generation of stimuli. Due to missing prior studies on pause detection in spontaneous speech it was unclear how to generate useful

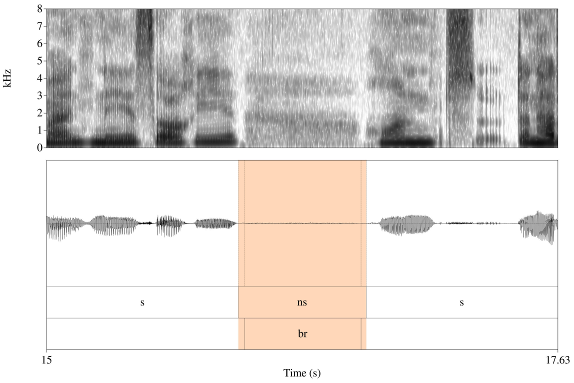

Figure 1 – Breath pause in a fluent section consisting of a breath noise surrounded by two very short silences.

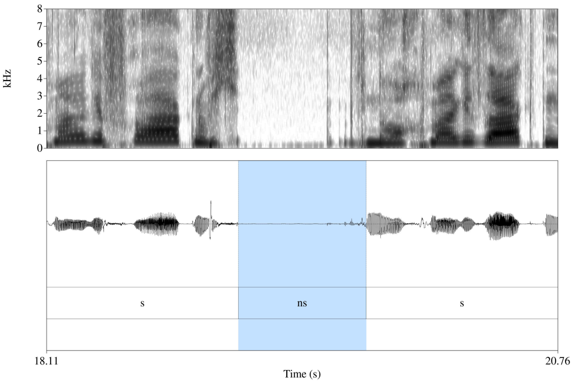

Figure 2 – Hesitation pause consisting of silence.

stimuli: How should speech samples be selected? How could they be manipulated, e.g., by replacing breath noises with silence or by removing silence?

A side interest of this exploratory study refers to the question: How well do automatic procedures detect speech pauses? Speech pause detection as voice activity detection has been widely used in automatic speech recognition since at least 1975 [6]. They are also useful in end-of-turn detectors in spoken dialogues systems [7] and for the automatic detection of filled pauses [8]. From a perspective of *human* pause detection it would be interesting to see the differences to automatic pause detection.

2 Experiment

2.1 Data selection

The data source for spontaneous speech was the GECO corpus [9] with dyadic conversations of German speakers (students unknown to each other with a free choice of topics). The channels for each speaker of a dyad was recorded separately with the consequence that there is no acoustic overlapping of both speakers. From sections with a participant in her speaker role, smaller samples were *randomly* selected as a fundament for the stimulus material. No pre-selection or prioritizing of certain types of pauses took place. Thus, there is no guarantee that the mix of pauses can be regarded as balanced. However, there was a substantial amount of pauses with breath noises and pauses that show some type of hesitation. Illustrations of different types of pauses with some articulated context can be found in Fig. 1-4.

2.1.1 Types of pauses

The random selection of 160 seconds of total speaking time in the samples contained 52 pauses. The great majority consisted of pauses with breath noises (n=32) which appears in fluent sections, i.e., without any disfluency like a hesitation. On average these pauses had a duration of 843 ms. Parts of this duration contained silence and one part contained the breath noise (on average 471 ms). There were also a few pauses in fluent sections without breath noises (n=3) with a mean duration of 696 ms.

A special category are pauses that contained laughter (n=3) which can also be regarded as breath noises. Those pauses were considerably longer (mean: 1563 ms).

The last category are pauses with a hesitation (n=14). The type of hesitation varied between those with just silence (most of them), with a lengthening and a silent portion, or a creaky voice in the prepausal articulation followed by silence. There were only two cases with a filler particle

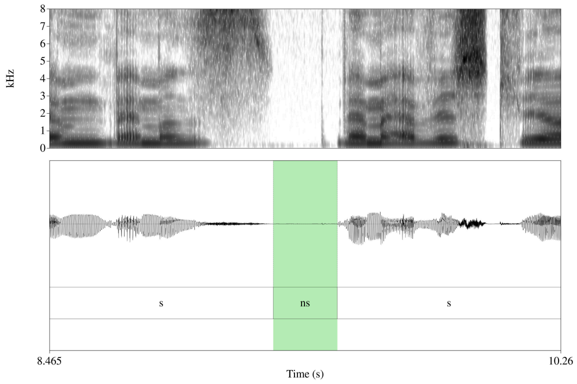

Figure 3 – Hesitation pause with a lengthened articulation before the silence.

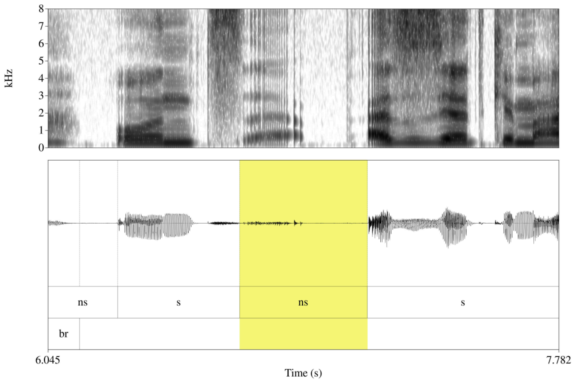

Figure 4 – Hesitation pause with a filler particle and two very short silences before and after.

(here "uhm") which is often considered as the typical hesitation pause. None of the pauses with a hesitation included a breath noise. Only two of them could be regarded as part of a false start that has been corrected (repair). Hesitation pauses were on average 455 ms long and thus considerably shorter than breath pauses.

2.2 Manipulations and stimuli

The stimuli had varying durations: 5, 10, 15, and 50 sec. For each of these four classes two stimuli were used in their original version and two stimuli in their manipulated version leading to a total of 16 stimuli.

Each selected sample file was copied to generate another file with manipulated pauses. Two types of manipulation were applied: In pauses with a breath noise the breath noise was replaced with silence. The example in Fig. 5 shows a pause of 800 ms containing a breath noise of 400 ms in the original version. Its manipulation shows the same 800 ms pause but without any breath noise. In pauses without a breath noise, i.e., all pauses with a hesitation, the silence was removed (actually set to 1 ms). There was no silence but there were other pause cues like lengthening or creaky voice. The missing coarticulation at the location of the cut could be a cue and also a false start, if present, could still be perceived. In pauses with filler particles the filler particle was removed as well. The example in Fig. 6 shows a hesitation pause with a silent portion in the original version and no silence in the manipulated version.

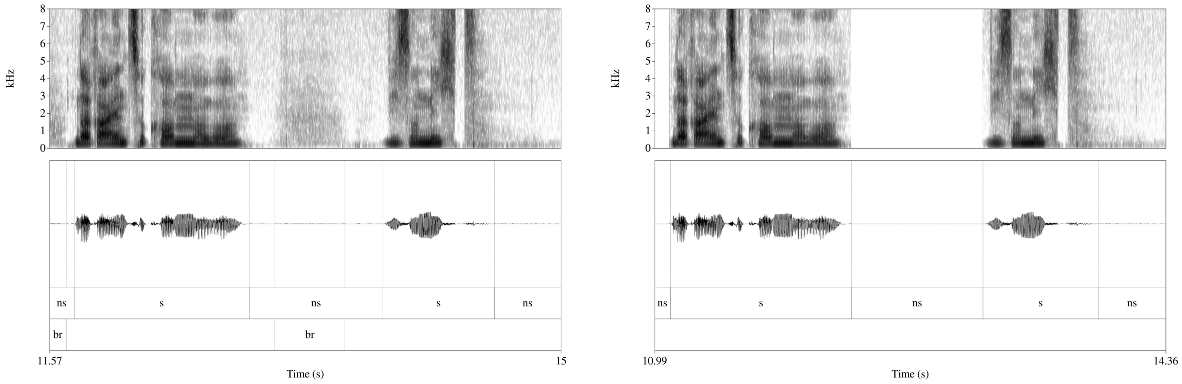

Figure 5 – Original (left) and manipulated version (right): in the manipulated version, the pause including the breath noise was replaced with silence.

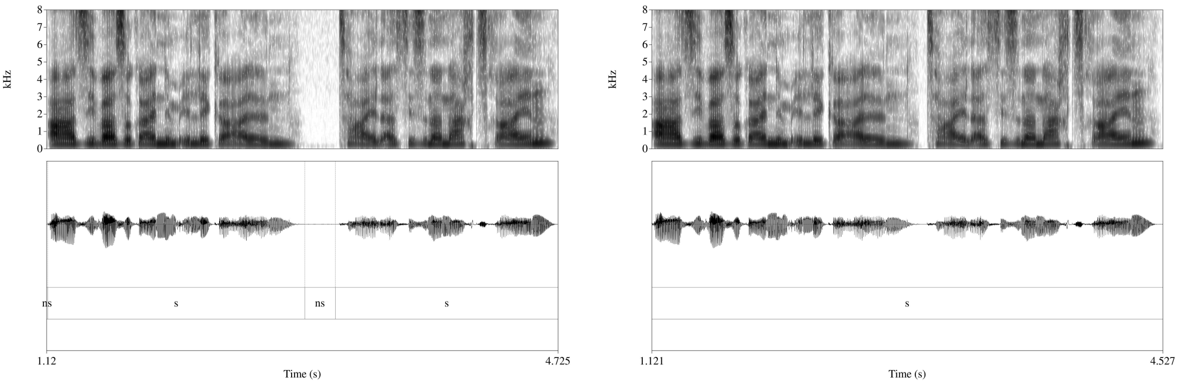

Figure 6 – Original (left) and manipulated version (right): in the manipulated version, the pause following a lengthened syllable was removed (the remaining silence comes from a plosive's closure phase).

2.3 Detection test

Twelve students from an introductory phonetics class served as subjects. They all had a basic knowledge of Praat including some skills concerning the annotation of speech signals in Praat with TextGrids. They were sent the audio files as wav-files in a randomised order (named 01-16.wav) and a prepared Textgrid with an empty point tier for each audio file.

In the instruction it was made clear that the window of the wave form and possibly the spectrogram should be minimised as much as possible in order to prevent any visual information for the students when the cursor is running over the speech signal including some presumed pauses. For this reason we asked the subjects to either close the eyes or to look away from the screen. While listening to an audio file with the prepared TextGrid on a point tier in Praat, participants were asked to tap the enter key as soon as they detected a pause which would then create a point in the TextGrid.

2.4 Segmentation

An experienced annotator for pauses segmented and labelled the pauses as stretches of one of the following categories: silence, breath noise, speech.

In addition, a widely used automatic pause detection routine [10] was applied for reasons of comparison as well. We let the script segment the speech with two different settings: one with a pause threshold of 50 ms and the other with a 200 ms threshold (silence threshold was kept at the default of -25 dB for both).

For the analysis of the reaction time of the subjects the end of an inter-pause unit (i.e., the beginning of a pause) was taken as a reference point. A negative value reflects a detection up to 500 ms before the beginning of the silence, a positive value refers to a detection after the beginning of the silence (maximally 1000 ms).

3 Results

3.1 Differences between subjects

In Table 1 the results for the detection rate and the reaction timing of the individual subjects are illustrated. Interestingly, there is no subject who detected all 89 pauses occurring with silences in the stimuli (the manipulated pauses with no silence not considered here).

As expected there are differences between the subjects regarding their detection rate and how fast they could detect the pauses. Subjects 2 and 5 are amongst the group with the slowest reaction times and also missed relatively many pauses. The opposite can be seen for subjects 4 and 6. They are the fastest subjects and missed only a few pauses. But there are also subjects who show other combinations of speed and detection rate like subject 12 who was rather slow but showed a good detection rate.

Table 1 – Average reaction times (in ms), number of detected and missed pauses, number of detections with negative values (before the beginning of the pause in the speech signal) for the individual subjects.

Subject	**Reaction time**	**#detected**	**#missed**	**#negative value**
01	233	74	15	1
02	412	53	36	4
03	180	56	33	9
04	110	78	11	17
05	371	48	41	1
06	165	78	11	16
07	146	68	21	10
08	231	71	18	1
09	155	58	31	6
10	225	78	11	3
11	261	83	6	0
12	334	79	10	0

3.2 Differences between pause types

3.2.1 Pauses with breath noises

Pauses in the condition with breath noises in the original version were generally detected by all subjects. There are only a few cases where such a pause was missed by a single subject. This finding holds for the original, as well as for the manipulated versions without any breath noises. The automatic detection was unproblematic for this condition, the only breath noise that was picked up as not part of a pause was a voiced inhalation that was part of laughing.

3.2.2 Pauses with hesitations

The detection rate of the hesitation pauses in the original versions strongly varied between all subjects. In the manipulated versions, without any silence, the detection rate was always lower than in the original version containing the silences. However, despite the complete absence of silence there was always at least one subject who detected a pause. The automatic detection has some problems here, as filler particles are picked up as sound segments (which they are of course). This concerns mainly voiced filler particles but occasionally includes glottal fillers.

3.2.3 Pauses without silences

The manipulated versions where silences were removed also evoked responses. In about 25% of those cases a pause was detected in this condition. That means that in a quarter of all cases a pause was perceived although there was no silence present.

It should be noted here that there were four cases where neither of the automatic nor the human annotator inserted a pause but a couple of participants (between 4 and 6) detected a pause. This was generally the case after a lengthened syllable (sometimes creaky) or a filled pause within fluent speech. This is probably also linked to the length of the respective speech segment as speakers try to predict the next pause. Here, syntax and utterance end projection play a role, as can be illustrated in the following example uttered without any silence: "... der war ganz ordentlich angezogen | eigentlich | und dann hab ich halt gemeint ..." (... he was properly dressed | actually | and then I said ...) where | marks a location of a syntactic boundary with a *probable* or expected pause. The automatic detection was impossible for this condition.

3.2.4 Automatic pause detection

The automatic pause detection used here did a relatively good job segmenting speech into silent and sounding parts. Both differed from human annotation in how they dealt with laughter and filled pauses. The two automatic versions have problems with frication noise preceding or following a pause and often assign it to the pause. The version with the 200 ms silence threshold is less good at finding shorter pauses but more robust against interpreting less loud speech segments as pauses: The 50 ms threshold version frequently falsely inserts a pause when there is a fricative, affricate, or aspiration (when in proximity of a pause both automatic versions struggle with this), a longer closure phase from a plosive, or syllables that are pronounced less loud (and often include frication noise).

4 Discussion

4.1 Human detection of pauses

No subject was able to detect all pauses. This finding suggests that the detection of speech pauses is not as straightforward as it might sound. The abilities of the individuals with regard to pause detection can be quantified using detection rate and reaction time. As expected, there were larger differences between the individuals. The skill of fast pause detection combined with an end-of-utterance projection can play a central role for turn-taking in conversations. There, the end-of-turn projection is needed for a timely place to take the turn [11] but also to give some feedback in form of a backchannel utterance without taking the turn [12].

The detection of pauses with breath noises was generally not a problem, which holds for all subjects, even if the breath noises were silenced. The comparably long durations in those pauses were long enough also for the slower subjects. The slowest subject had a mean reaction time of 412 ms which still fits in the mean pause time of 843 ms for the breath pauses but also in the mean pause time of 455 ms for the hesitation pauses. The fastest subject with an average reaction time of only 110 ms (and various negative values) does probably not need a longer pause duration for its detection.

An important observation is that perceived pauses do not necessarily require overt silence. If silence is not needed to detect pauses (at least for some subjects), then we should make a distinction between pauses for speech production and pauses for speech perception.

This suggested distinction is enforced by the observation that silences located after lengthened hesitations and filler particles do not lead to the same pause detection patterns found after fluent speech sections. Both findings reinforce our doubts that pauses should be clearly divided into only two categories, namely *silent pauses* or *filled pauses*.

4.2 Automated detection of pauses

In general, the automated detection of pauses applied here worked fine for pauses with silences and breath noises when the threshold was low enough to ignore breath noises.

The exact segmentation of the beginning and the end of these pauses is always problematic when an inter-pause unit ends with a fricative or a stop (including its aspiration). However, segmentation and detection are two different tasks and should be considered separately here.

The detection of pauses without any silence was not possible in the automatic way, as expected. In contrast some of the subjects were able to perceive pauses also without any silence. In this respect automatic and human detection abilities differ which might give rise to thinking about additional techniques to make automatic detection of pauses more human-like.

5 Summary and Conclusion

In this study, we compared different cues for human and machine pause detection (silence, breath noise, without silence but with prosodic/syntactic cues before the pause). We found that different types of pauses are dealt with in different ways by human and automated pause detection. While both are very good at detecting pauses that include breath noises or stretches of silence, they differ when dealing with no silence. Whereas some of the participants did detect pauses without any silence by relying on prosodic/syntactic cues, the automated detection generally ignored them. The same is the case when pauses were filled with laughter or filler particles. The difference arises from the script looking at (vocal) activity only, while human subjects have a broader definition of what qualifies as a pause. However, for both groups silence is the strongest indicator of a perceived pause.

This small-scale study was set to explore several aspects: One goal was to test whether this type of remote test is feasible and leads to valid results. In our view, the test procedure worked without serious problems. The same holds for the results. One clear constraint is that basic skills of using Praat and working with TextGrids are necessary to perform the test. This restricts the number of possible subjects dramatically. Another critical point is that students in a speech science discipline, even at a beginner level, may have a more differentiated view on pauses than other people. A critical technical aspect is that it is out of control of the experimenter whether the subjects get some visual information from the display of the speech signal despite the instruction being very clear on this point. Here the experimenters have to rely on a trustful performance of tasks by their subjects (which hopefully reflects the trustful relationship to their students in this case).

As far as we can see, the variation of stimulus length did not have a negative effect on the test performance. Maybe a warm up with two or three stimuli could be performed first in the future.

For this study two ways of manipulation were applied: on the one hand replacing breath noise with silence (for the long pauses in fluent sections) and removing silence and possible filler particles (for pauses in disfluent sections). Both manipulations did not lead to the perception of unnaturalness and obvious manipulation because stretches of coarticulation remained untouched. Further manipulations can test different durations of silences in longer pauses, e.g., in steps of 100 ms.

Another interesting direction of research would include the specific selection of samples with a high vs. low level of expectancy of upcoming pauses based on syntax, possibly with manipulations of the pause.

Besides that a large, rather unexplored field is the perception and the detection of pauses with filler particles. The term ’filled pause’ is commonly vaguely defined (or not defined at all) for speech *production*. It usually remains unclear whether this term is restricted only to the filler particle or to the filler particle plus the surrounding silences. The *perceptual* state of this term is not clear as well. Are ’filled pauses’ perceived as *pauses* anyway? Filler particles very often occur with other elements of hesitation including silence (but not necessarily so) and lengthening. What are the effects of removing only the filler particles in different combinations of hesitation?

We see options to refine and extend automatic pause detection towards human pause perception. This concerns automated detection of 'silent' pauses, of pauses with various types of filler particles (not only *uh* and *uhm*), and pauses without any silences.

Further studies could profit from including other conditions such as pauses in positions where syntax or prosody is not useful for pause prediction. It would also be of interest to include more pauses that contain no silence at all but only breath noise or a filler particle.

References

[1] DUEZ, D.: *Acoustic correlates of subjective pauses. Journal of Psycholinguistic Research*, 22(1), pp. 21–39, 1993. doi:10.1007/BF01068155.

[2] STRANGERT, E.: *Speaking style and pausing. Phonum*, 2, pp. 121–137, 1993.

[3] SWERTS, M.: *Filled pauses as markers of discourse structure. Journal of Pragmatics*, 30(4), pp. 485–496, 1998. doi:10.1016/s0378-2166(98)00014-9.

[4] CARLSON, R., J. HIRSCHBERG, and M. SWERTS: *Cues to upcoming Swedish prosodic boundaries: Subjective judgment studies and acoustic correlates. Speech Communication*, 46(3-4), pp. 326–333, 2005. doi:10.1016/j.specom.2005.02.013.

[5] BOERSMA, P. and D. WEENINK: *Praat: doing phonetics by computer*. 2019. URL `http://www.praat.org/`.

[6] RABINER, L. R. and M. R. SAMBUR: *An algorithm for determining the endpoints of isolated utterances. Bell System Technical Journal*, 54(2), pp. 297–315, 1975. doi:https://doi.org/10.1002/j.1538-7305.1975.tb02840.x.

[7] MICHAEL, T. and S. MÖLLER: *Retico: An open-source framework for modeling real-time conversations in spoken dialogue systems*. In P. BIRKHOLZ and S. STONE (eds.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 134–140. TUDpress, Dresden, 2019.

[8] REICHEL, U. D., B. WEISS, and T. MICHAEL: *Filled pause detection by prosodic discontinuity features*. In P. BIRKHOLZ and S. STONE (eds.), *Studientexte zur Sprachkommunikation: Elektronische Sprachsignalverarbeitung 2019*, pp. 272–279. TUDpress, Dresden, 2019.

[9] SCHWEITZER, A. and N. LEWANDOWSKI: *Convergence of articulation rate in spontaneous speech*. In *Proceedings INTERSPEECH*, pp. 525–529. 2013.

[10] DE JONG, N. H. and T. WEMPE: *Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods*, 41(2), pp. 385–390, 2009. doi:10.3758/BRM.41.2.385.

[11] HELDNER, M. and J. EDLUND: *Pauses, gaps and overlaps in conversations. Journal of Phonetics*, 38(4), pp. 555 – 568, 2010. doi:https://doi.org/10.1016/j.wocn.2010.08.002.

[12] TRUONG, K., R. POPPE, I. DE KOK, and D. HEYLEN: *A multimodal analysis of vocal and visual backchannels in spontaneous dialogs*. In *Proceedings INTERSPEECH*, pp. 2973–2976. 2011.