ARE ‘SILENT’ PAUSES ALWAYS SILENT?

Malte Belz[1], Jürgen Trouvain[2]

[1]Humboldt-Universität zu Berlin, [2]Universität des Saarlandes
malte.belz@hu-berlin.de

ABSTRACT

This study explores the phonetic activity in speech pauses. The often used term ‘silent pause’ (as opposed to ‘filled pause’) implies that these pauses are exclusively made up of silence. However, there is evidence that most pauses contain phonetic particles such as breath noises or tongue clicks. The investigated samples of two speaking styles (radio news vs. spontaneous dialogues) demonstrate that only a minority of speech pauses are completely silent. In addition, the clear distinction between silent and non-silent pause phases allows for a better analysis and understanding of phonetic particles correlated to respiratory, articulatory or physiological activity. In this vein, we give a detailed description of the annotation of phonetic particles and their challenges, followed by an exemplary analysis of the most frequent pause pattern.

Keywords: silent pauses; filled pauses; non-verbal vocalisations; breath noises

1. INTRODUCTION

This paper aims at a critical consideration of the technical term ‘silent pause’ by exploring phonetic corpora. The notion of ‘silent pause’ (or ‘unfilled pause’) is frequently used in contrast to so-called ‘filled pauses’ [11, 6, 2]. The latter term is often used to describe fillers such as [ə] or [əm]. In contrast, “silent pauses” would be defined as pauses *not* containing any filler particles. However, it remains unclear whether a ‘silent pause’ is completely silent in a phonetic sense – which is implied by the term – or whether it can contain subtle phonetic particles such as breathing noises, tongue clicks or other unspecified articulatory activity.

Silence in a narrow acoustic sense is probably seldom observable in data used in phonetics and linguistics. From a phonetic point of view, silence can be investigated in different ways. Acoustically, it equals the absence of any correlates of vocalisation. Perceptually, the listener interprets the acoustic signal as silent although it might exhibit subtle events. In addition, silence is often used as a synonym to pause and/or to prosodic phrase break. However, we define silence not necessarily as the entire pause but as a *phase* within a perceived pause, more concretely a phase that does not contain any phonetic particles or events, such as breathing, clicking, etc. To avoid confusion, we use the term *pause* in this study as the acoustic correlate of a *perceived* pause (cf. [12]) within a stretch of speech. Such a pause can contain silent phases and other, unspecified phonetic particles within stretches of speech.

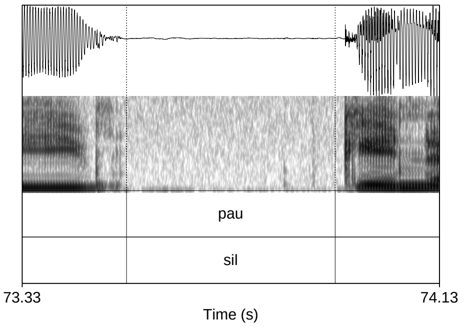

Figure 1: Example of a completely silent pause (400 ms), taken from speaker K in the investigated GECO corpus [15].

If a pause consists entirely of silence, it would be a completely silent pause (see Figure 1). If a pause contains phonetic particles, it would not be termed as silent pause (though it can contain silent phases, see Figure 2).

To find silent phases and phonetic particles in perceived pauses, they should always be determined with the help of the acoustic signal. Thus, we exclude perceived pauses purely based on cues at edges of prosodic phrases such as final lengthening, final lowering or initial strengthening.

The aims of this paper are threefold. First, we would like to demonstrate that ‘silent pauses’ are often not silent in a phonetic sense, as described above. Second, we explore the types of phonetic particles included in perceived pauses as well as their relative frequencies. We use two samples of opposed speaking styles, namely radio news speech produced by professional speakers and spontaneous dialogues between students. Third, we describe the annotation procedure and challenges of marking phonetic parti-

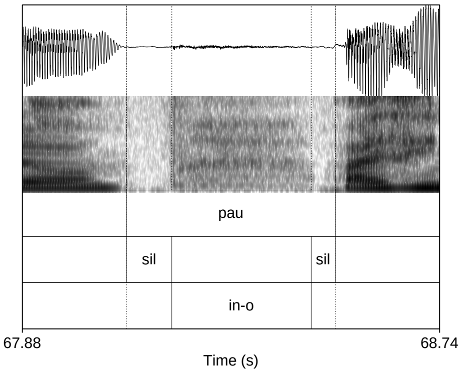

Figure 2: Example of a breath pause, taken from speaker K in the investigated GECO corpus [15]. The oral inhalation phase (289 ms) is sandwiched between short periods of silence (93 ms and 50 ms, respectively).

cles in perceived pauses.

2. METHOD AND DATA

We use a sample of the GECO corpus version 1.1 [15] and the DIRNDL corpus [5]. Thus, we contrast two very different registers, namely dialogic task-free spontaneous speech versus monologic scripted read speech in German.

We sampled the first three minutes of three dialogues (six German female speakers) of GECO as well as six German radio news items of DIRNDL. The actual articulation time of all speakers amounts to about 9 minutes in dialogues and about 4 minutes in radio news. The corpus is annotated in Praat [3] and queried with *emuR* version 1.1.1 [20].

2.1. Annotation

We use intervals in Praat on six tiers:

- speaking phases (TURN)
- pauses (PAUSE)
- silent phases (SILENCE)
- respiratory activity (RESP)
- articulatory activity (ARTIC)
- residual phenomena (ELSE)

On the TURN tier, we mark sections of continued speech production (tag: *spk*) that consist of more than mere backchanneling entities and which are not interrupted by the interlocutor, except for a short feedback utterance (tag: *fb*). We define feedback utterances as short lexical or non-lexical productions with potentially many functions, such as backchanneling [9], agreements, rejections, or comments. Typical examples for lexical feedback utterances are *ja okay*, for non-lexical ones a sequence often transcribed as *mhm* or *hm*. In line with Schmidt [14], we consider *hm* as a neutral consonant with a closed mouth and with only intonation as the carrier of phonetic information. These items are free of segmental-lexical and grammatical information. Often, feedback utterances with *hm* contain a bisyllabic *hm-hm*. To the left of a speaking phase, we annotate an antecedent phase (tag: *ante*), if there is any activity different from silence. For practical reasons, we delimitate this phase to a maximum of 500 ms. To the right of a speaking phase, we annotate a follow-up phase, respectively (tag: *post*). The maximum follow-up phase we marked was 676 ms long.

On the PAUSE tier, we mark perceived pauses that occur within *spk* phases with the tag *pau*. We define *pau* as an audible interruption of the articulatory flow of words within a speaking phase. We exclude closure phases of stops (right after a pause a default value of 50 ms was considered as closure phase of the stop) as well as prolonged closure phases in slips of the tongue from this definition. Additionally, we annotate fillers on this tier. We define them as being always a part of *spk*. In general, the annotation of fillers is handled rather idiosyncratically, compare [2, 6, 7], among many others. In fact, their highly variable acoustic realization would require a precise phonetic annotation, which goes beyond the purpose of this study. Here, we will confine ourselves to two categories, a vowel-only filler, e.g., *äh* (tag: *f-v*), and a vowel-nasal filler, e.g., *ähm* (tag: *f-vn*). Both filler categories can potentially be preceded by glottal stops. Other particles presumably functioning as fillers, for example clusters of stops and fricatives (*pff*) or sequences of glottal stops [1], are marked with *f-o*.

On the SILENCE tier, we mark those phases within *pau* with *sil* which do not contain any other phonetic particles.

On the RESP tier, we mark breathing activity within *ante*, *spk*, and *post* phases with the tags *in-o*, *in-n*, *ex-o* and *ex-n* for oral and nasal inhalation and exhalation, respectively.

On the ARTIC tier, we mark every non-lexical articulatory event. Among them, clicks, either as singletons or as sequences, are typical examples (tag: *cl*). Unclear but nevertheless clearly audible fragments are marked with an *x*.

Finally, we mark various phenomena such as laughing and vegetative sounds such as coughing, throat-clearing, and swallowing on a residual tier called ELSE. Unclear cases can be marked with *x* on every tier.

A special case is laughing, as it is not a part of the articulatory flow as described for the speaking phase. It can have some marks on the respirational tier occurring synchronously with speaking. Laughter can potentially occur both within and between tages on the TURN tier. Laughter is included in *pau*, as there is not much articulatory activity involved.

2.2. Annotation challenges

On the TURN level, we face the problem to concretely determine and separate speaking phases and feedback utterances. There is no standard definition of feedback utterance, which can lead to problems as exemplified by the following three observations. A confirmation request consisting of a repeated word with a rising intonation could be counted both as a speaking phase or a feedback utterance. There is a possibility but not a need to consider an isolated laugh as a feedback utterance. Third, some feedback utterances serve both the function of giving feedback and beginning a turn [4].

We decided to set the maximum extension of the antecedent phase to the arbitrary value of 500 ms. However, there are examples where the antecedent clearly exceeds this threshold.

On the PAUSE level, we often see intervals between a feedback utterance and a speaking phase that extend to several seconds, but there are also intervals shorter than a second. These intervals are not regarded as pauses here but could, in theory, also be considered as pauses between stretches of speech production. In this paper, we focus on intra-turn pauses, as extra-turn pauses could be seen as a different phenomenon.

On the SILENCE level, the term *silence* is problematic, too. A silent phase, as defined here, can contain some noises visible in the signal and often (but not always) audible. This can be due to extraneous non-vocal noises (e.g., door slamming) but also due to unspecified noise production presumably produced by the speaker (e.g., head scratching) but also due to unknown origin. Superficially, such a picture contradicts the idea of silence.

As to the RESP level, most cases of breath noises can be audibly determined (with oral inhalation as a frequent and also the most salient sub-type). However, the clarity of categorization is not possible for all sub-types of respiration, as to our experience with the data at hand.

Clicks show a great range of variability [18]. They can occur as one or as a series of multiple pulses, which makes the distinction between one or several events hard to determine. Often, their intensity is rather low, which makes it difficult to decide whether they should be annotated or not. They are often clearly visible in the spectrogram, but not always audible when a larger context (> 2 sec) is taken into account. In the latter cases, they are not chosen for annotation.

During the annotation, we noticed that there are instances of presumably glottal activity before or within inhalation. Where these instances were audible involving the larger context, they are annotated with *g*. Other unknown sounds occured on the ARTIC tier, where the acoustic signal did not suffice to determine a clear phonetic category.

3. RESULTS

3.1. Completely silent vs. partially silent pauses

Referring to the question we ask in the title, Figure 3 shows that in the news samples only 16 % of all pauses are completely silent. In the dialogues, the relative number of completely silent pauses is only about a third (34 %). Thus, phases that have been referred to as silent pauses are in many cases phonetically not silent.

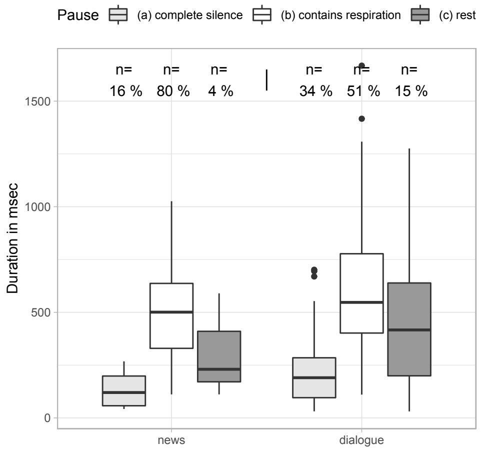

Figure 3: Durations of pauses which (a) are completely silent, (b) contain respiratory noises, and (c) the rest (i.e. pauses which are neither (a) nor (b)). The relative number of pause types per speech style is given in percent above each box.

Confirming other studies [8, 17], completely silent pauses are shorter than partially silent pauses, as can be seen for both styles in Figure 3. News show shorter pause duration means than dialogues for complete silences (134 ms vs. 241 ms), as well as for partially pauses that are not completely silent (523 ms vs. 602 ms, respectively).

When analysing the total duration of all pauses and silent phases, we see that in the dialogues 50% of the total pause duration (65.067 s) consists of silent phases (34.567 s). In the news, the proportion of silent phases is only 35% (11.686 sec out of

29.356 s total pause duration). A major part of pause duration is used for the production of non-silent phonetic activity.

3.2. Types of particles

Table 1 shows that many particle types do not exist in the news style, for example feedback utterance, fillers or laughs. Although this is not a surprising result, it confirms the expectations to both investigated styles. It can be seen as a qualification of the dichotomies of dialogic spontaneous vs. monologic read speech. On the ELSE level, we found no affective sounds and only few vegetative sounds, such as swallowing or throat clearing. In contrast, laughter occurred several times, which confirms former analyses of conversational corpora [19].

Table 1: Counts of annotation tags on the five tiers for dialogues and radio news and normalized counts per minute (cpm).

		dialogue		news		
		N	cpm	N	cpm	total
TURN	fb	96	10.67	0	0	96
	spk	81	9	6	1.5	87
	ante	58	6.44	1	0.25	59
	post	23	2.56	0	0	23
	post/ante	5	0.56	0	0	5
PAU	pau	144	16	67	16.75	211
	f(illers)	29	3.22	0	0	29
RESP	in	133	14.78	56	14	189
	ex	27	3	0	0	27
	unclear	3	0.34	0	0	3
ARTIC	cl	32	3.56	16	4	48
	g(lottal)	16	1.78	0	0	16
	else	13	3	0.34	16	4
ELSE	laugh	22	2.44	0	0	22
	other	8	0.89	0	0	8

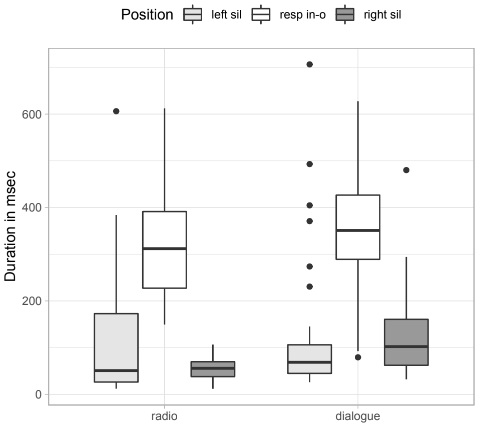

Figure 4: Durations of the pattern *silence – oral inhalation – silence* (*left sil – resp in-o – right sil*) in dialogues (n = 37) and news (n = 34).

3.3. Preparation in pauses and in antecedent phases

Roughly half of all speaking phases show an antecedent phase, nearly all of which are characterised by a respiratory noise, which is regularly an oral inhalation.

The most frequent pause pattern that consists of more than two entities adjacent to each other is the sequence of silence plus oral inhalation plus silence. An example is depicted in Figure 2, and the general pattern is quantified in Figure 4. There, we clearly see that the breath duration is by far longer than the two silent phases, which is in line with the results given above.

4. DISCUSSION AND CONCLUSION

A clear shortcoming of this study is the small sample analysed, which does not allow for any generalization. However, the challenges of categorizing and annotating a continuous and highly subjective phenomenon remain the same. The aim of this paper was, therefore, to show the need for a thoroughly discussed annotation scheme for phonetic particles before attending to studies of a larger scale.

As has been shown in detail, most pauses do not contain nothing (though a subpart are phonetically really silent pauses), but contain various phonetic particles. The most frequent items are breath noises, especially oral inhalation. There is also a considerable amount of tongue clicks and laughs, but also events due to presumably glottal activity.

It seems that systematic patterns are not only produced using combinations of fillers and pauses [10], but also within this domain of subtle phonetic particles. Whether these patterns turn out to be context-dependent or idiosyncratic is open to further research considering also a combination of acoustic, respiratory and articulatory evaluation [13, 16]. Likewise, such physiological measurements in combination with acoustic recordings would be needed in order to find out more about the phonetic processes of articulatory preparation and their acoustic correlates in speech.

Annotation of non-vocal verbalisations seems to be a rather straightforward task for professional read speech, whereas the annotation reveals difficulties at various levels for spontaneous speech. A concrete determination of speaking phases (or turns) and feedback utterances requires a conceptually coherent and theoretically solid concept to be applied to the acoustic signal, which is often not the case. Additionally, there is a mismatch between the visual inspection of these phonetic particles – which can be quite subtle – and their auditory perception.

5. REFERENCES

[1] Belz, M. 2017. Glottal filled pauses in German. Eklund, R., Rose, R. L., (eds), *Proceedings of DiSS 2017* number 58(1) in TMH-QPSR 5–8.

[2] Belz, M., Sauer, S., Lüdeling, A., Mooshammer, C. 2017. Fluently disfluent? Pauses and repairs of advanced learners and native speakers of German. *International Journal of Learner Corpus Research* 3, 118–148.

[3] Boersma, P. 2001. Praat, a system for doing phonetics by computer. *Glot International* 5, 341–345.

[4] Drummond, K., Hopper, R. 1993. Back Channels Revisited. *Research on Language & Social Interaction* 26, 157–177.

[5] Eckart, K., Riester, A., Schweitzer, K. 2012. A Discourse Information Radio News Database for Linguistic Analysis. In: Chiarcos, C., Nordhoff, S., Hellmann, S., (eds), *Linked Data in Linguistics*. Springer 65–75.

[6] Eklund, R. 2004. *Disfluency in Swedish human-human and human-machine travel booking dialogues*. Dissertation.

[7] Gósy, M., Gyarmathy, D., Beke, A. 2017. Phonetic analysis of filled pauses based on a Hungarian-English learner corpus. *International Journal of Learner Corpus Research* 3, 149–174.

[8] Grosjean, F., Collins, M. 1979. Breathing, pausing, and reading. *Phonetica* 36, 98–114.

[9] Iwasaki, S. 1997. The Northridge earthquake conversations: The floor structure and the 'loop' sequence in Japanese conversation. *Journal of Pragmatics* 28, 661–693.

[10] de Leeuw, E. 2007. Hesitation Markers in English, German, and Dutch. *Journal of Germanic Linguistics* 19, 85–114.

[11] Lickley, R. J. 2015. Fluency and Disfluency. In: Redford, M. A., (ed), *The Handbook of Speech Production*. John Wiley & Sons, Inc 445–469.

[12] Nakatani, C. H., Hirschberg, J. 1994. A corpus-based study of repair cues in spontaneous speech. *Journal of the Acoustical Society of America* 95, 1603–1616.

[13] Rasskazova, O., Mooshammer, C., Fuchs, S. 2018. Articulatory settings during inter-speech pauses. Belz, M., Mooshammer, C., Fuchs, S., Jannedy, S., Rasskazova, O., Żygis, M., (eds), *Proceedings of the Conference on Phonetics & Phonology in German-speaking countries (P&P 13)* 161–164.

[14] Schmidt, J. E. 2001. Bausteine der Intonation? *Germanistische Linguistik* 157-158, 9–32.

[15] Schweitzer, A., Lewandowski, N. 2013. Convergence of Articulation Rate in Spontaneous Speech. *Proceedings of Interspeech* 525–529.

[16] Scobbie, J. M., Schaeffler, S., Mennen, I. 2011. Audible Aspects of Speech Preparation. *ICPhS* 1782–1785.

[17] Trouvain, J., Fauth, C., Möbius, B. 2016. Breath and Non-breath Pauses in Fluent and Disfluent Phases of German and French L1 and L2 Read Speech. *Proceedings of Speech Prosody (SP8)*. Boston 31–35.

[18] Trouvain, J., Malisz, Z. 2016. Inter-Speech Clicks in an Interspeech Keynote. *Interspeech* 1397–1401.

[19] Trouvain, J., Truong, K. P. 2012. Comparing Non-Verbal Vocalisations in Conversational Speech Corpora. *4th International Workshop on Corpora for Research on Emotion Sentiment & Social Signals* 36–39.

[20] Winkelmann, R., Jaensch, K., Cassidy, S., Harrington, J. 2018. emuR. R package version 1.1.1.