The following paper was presented at The 9th Workshop on Disfluency in Spontaneous Speech (DiSS 2019) held at ELTE Eötvös Loránd University in Budapest, Hungary on 12–13 September, 2019.

Title: Fill the silence! Basics for modeling hesitation

Author(s): Simon Betz and Loulou Kosmala

Abstract: In order to model hesitations for technical applications such as conversational speech synthesis, it is desirable to understand interactions between individual hesitation markers. In this study, we explore two markers that have been subject to many discussions: silences and fillers. While it is generally acknowledged that fillers occur in two distinct forms, *um* and *uh*, it is not agreed on whether these forms systematically influence the length of associated silences. This notion will be investigated on a small dataset of English spontaneous speech data, and the measure of distance between filler and silence will be introduced to the analyses. Results suggest that filler type influences associated silence duration systematically and that silences tend to gravitate towards fillers in utterances, exhibiting systematically lower duration when preceding them. These results provide valuable insights for improving existing hesitation models.

DOI: https://doi.org/10.21862/diss-09-004-betz-kosm

Citation (JIPA): Betz, Simon & Loulou Kosmala. 2019. Fill the silence! Basics for modeling hesitation. In: R. L. Rose & R. Eklund (eds.), *Proceedings of DiSS 2019, The 9th Workshop on Disfluency in Spontaneous Speech*, 12–13 September, 2019, Budapest, Hungary, 11–14.

The complete proceedings for DiSS 2019 are available as follows.

ISBN: 978-963-489-063-8

DOI: https://doi.org/10.21862/diss-09

DiSS 2019 was sponsored by The Faculty of Humanities, ELTE Eötvös Loránd University and the International Speech Communication Association (ISCA).

Fill the silence! Basics for modeling hesitation

Simon Betz [1] and Loulou Kosmala [2]
[1]Phonetics and Phonology Workgroup, Bielefeld University, Bielefeld, Germany
[2]PRISMES EA 4398/SeSyLiA, Sorbonne Nouvelle University, Paris, France

Abstract

In order to model hesitations for technical applications such as conversational speech synthesis, it is desirable to understand interactions between individual hesitation markers. In this study, we explore two markers that have been subject to many discussions: silences and fillers. While it is generally acknowledged that fillers occur in two distinct forms, um *and* uh*, it is not agreed on whether these forms systematically influence the length of associated silences. This notion will be investigated on a small dataset of English spontaneous speech data, and the measure of distance between filler and silence will be introduced to the analyses. Results suggest that filler type influences associated silence duration systematically and that silences tend to gravitate towards fillers in utterances, exhibiting systematically lower duration when preceding them. These results provide valuable insights for improving existing hesitation models.*

Introduction

When speakers are engaged in face-to-face interactions, their productions contain frequent hesitations. Hesitation can be defined as "the temporary suspension of flowing speech" (Lickley, 2015: 40). This paper focuses on the distribution and duration of two common hesitation markers: fillers and silences, especially in co-occurrence, i.e. silences that appear in the same utterances as fillers.

Fillers and silences are said to be the most frequent types of hesitations (Shriberg, 1994, Eklund, 2004). Their temporal features have been explored by numerous researchers, and some of them have labelled them as signals of suspension (Clark & Fox Tree, 2002). Clark and Fox Tree's main hypothesis is that fillers signal a speaker's intention to initiate a delay, and that um signals a major delay, while uh signals a minor one. This was in part explained by the fact that more pauses occurred with um than with uh. The presence of a pause can thus play a role in this signaling-a-delay hypothesis. Other studies have looked at the co-occurrence of fillers and silences: Grosjean and Deschamps (1972) found that fillers were often combined with silences both in French and English; Smith and Clark (1993) argued that um was followed by a longer pause than uh because speakers intentionally chose between uh and um to signal their word retrieval difficulties; in a study of pauses in deceptive speech, Benus et al. (2006) found that um was more likely to be followed by a silence than uh, and that silences were longer when they followed turn-initial um. In investigating the cluster of disfluencies, Kosmala and Morgenstern (2017) found two recurrent combinations: filler + silence, and lengthening + filler. Betz and Lopez Gambino (2016) also found that speakers engaged in a description task sometimes produced a filler after an initial silence, which allowed them to buy time before planning the description.

However, the idea of uh and um being consciously chosen by speakers to signal an upcoming delay is questionable. Finlayson and Corley (2012) argued that the fact that fillers tend to precede silences does not necessarily mean that they are intentionally chosen. O'Connell and Kowal (2005) rejected the signaling hypothesis and more specifically the status of uh and um as interjections, and Schegloff (2010: 71) argued that although fillers can be associated with delay, they do not "announce" a delay, but rather "embody" it.

In line with these issues, we further explore the co-occurrence of fillers and silences. Clusters of multiple markers have seldom been the focus of analysis, so our aim is to provide insights about the interplay of hesitation markers in order to model them for technical applications. We focus on two broad topics in this investigation. First, we test if the challenged assumption by Clark and Fox Tree—that silence duration varies as a function of filler type—can be confirmed. Second, we extend the analysis by measuring distance between silences and fillers to test whether fillers can influence duration of silences that are further remote than their direct vicinity.

Corpus and methods

The materials used for this study are taken from the FILM corpus (Kosmala & Morgenstern 2019) which is a collection of recordings between 16 native English speakers (aged 18–23) engaged in face-to-face dyadic interactions in the form of a film interview in familiar settings. The participants knew each other fairly well, and interacted in pairs. The interviewer asked a series of 10 questions about the film to the interviewee, and the latter was asked to answer the questions as spontaneously as possible.

https://doi.org/10.21862/diss-09-004-betz-kosm

The total duration of the corpus is approximately 71 minutes.

We investigated the co-occurrence of fillers and silences in the data. Following Clark and Fox Tree (2002), we distinguished two types of fillers, uh and um. These fillers differ on the phonetic surface in being either a centralized vowel (uh) or a centralized vowel with a nasal (um). As explained earlier, according to Clark and Fox Tree (2002), these two types are mutually exclusive and denote either a minor or major delay, which can be quantified by measuring adjacent silences. For this investigation, we thus measured the duration of fillers and associated silences (i.e. silences in the same utterance, either preceding or following), using the ELAN software. We were only interested in co-occurring hesitations so we only selected utterances that contained both fillers and silences, yielding 722 silences and 303 fillers in total. For silences, we measured the distance in words from the fillers. 0 denotes the first position after a filler, values > 0 subsequent positions. -1 denotes the last position before the filler, values < -1 greater distance before a filler.

We first aim to test the hypotheses stated by Clark and Fox Tree (2002):

- Is the duration of silences associated with um higher than those associated with uh?
- Does the duration of the filler correlate positively with the duration of the associated silence?

Furthermore, we explored the notion of distance:

- What is the average distance of silences in utterances where they co-occur with fillers?
- Does the distance between silence and filler influence silence duration?

Results

Silence duration after um and uh

The duration of silences is on average 155 ms higher when the silence occurs in an utterance with a um-type filler as opposed to uh-type filler. We fitted a linear mixed effects model with silence duration as the dependent variable and filler type as the independent variable. We included as random effects random slopes for speakers, random slopes for distance between silence and filler, and random slopes for position of the filler within the utterance. The difference is significant: $p = 0.0084$, $t = 2.661$, $DF = 210.57$, $SE = 56.56$. The speakers showed great variability. Model comparisons using analyses of variance between the full model and the reduced model without random slopes for speakers yielded significant results ($p = 0.019$). This is likely due to the fact that the amount of fillers produced varies strongly per speaker, which has been attested for this dataset (Kosmala & Morgenstern, 2019) and has frequently been observed with other data as well (e.g. Betz & Lopez Gambino, 2016). We conducted exploratory post-hoc tests to see if filler rate per speaker correlates with silence or filler duration, but found no such interaction.

This study is based on a small dataset of interview-style interaction, in which we expect a lot of turn-initial fillers, but the model comparisons suggested no influence of position on the results. We conducted additional t-tests on the mean duration of utterance-medial and utterance-initial fillers and associated silences. The general idea that silences are longer when they co-occur with um-type fillers was confirmed for medial position only, but not for initial position, where the same tendency was observed, but failed to reach significance ($p = 0.044$, $DF = 56.4$ $t = 2.06$ for medial position; $p = 0.21$, $DF = 42.9$, $t = 1.26$ for initial position). Additionally, silences associated with fillers of both types in utterance-initial position were longer than those in medial position, but not significantly (for um: $p = 0.14$, $DF = 83$, $t = 1.5$; for uh: $p = 0.12$, $DF = 45.21$, $t = 1.6$). This difference might be clearer when analyzed on a bigger dataset, as it would be conceivable that turn-initial, planning-related hesitations span a significantly longer time.

Correlation of filler and silence duration

Clark and Fox Tree (2002) split the two types of fillers into prolonged and not prolonged fillers, yielding four types (um, u:m, uh, u:h). They found that prolonged fillers were associated with longer silences. In our data, the types were not divided a priori between prolonged and not prolonged; we rather fit linear regression models to see if duration of a filler correlated with the duration of the associated silence.

Our data contains 303 fillers, 83 of which appeared with no silences. We conducted this analysis for both uh and um, confirming the findings of Clark and Fox Tree (2002): longer uh and um were associated with longer silences in the utterance ($p = 0.017$, $t = 2.4$, $SE = 0.04$ for um; $p = 0.009$, $t = 2.7$, $SE = 0.07$ for uh). This finding, however, has to be taken with caution. When considering silences immediately following or preceding the filler, only the duration of silences preceding um correlated with the duration of the filler. This might be an artifact of the small size of our dataset and is up for future research to verify.

Distance between silences and fillers

As Figure 1 shows, the longest silences are directly adjacent to fillers, either preceding (-1) or following

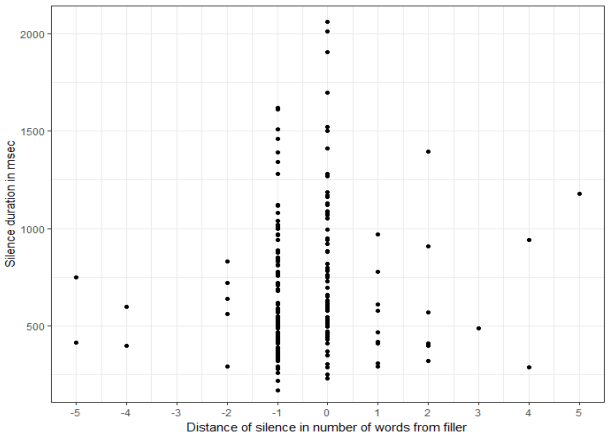

Figure 1. Silence duration and distance from fillers.

(0). The mean distance is lower for um (−0.6) than for uh (−0.5), but as a *t*-test reveals, this difference is not significant ($p = 0.7$, $DF = 83.7$, $t = -0.375$). However, the distance seems to influence the duration of silences. As can further be seen, silences before fillers (−1) are shorter than following fillers (0). This difference is significant and holds true for both um and uh type fillers ($p = 0.013$, $DF = 78.7$, $t = 2.5$ for um; $p = 0.017$, $DF = 40.4$ $t = 2.5$ for uh).

Discussion

Filler types

The claim that fillers are an intentional signal with a word status has been hotly debated in the research community. This study on a few hundred instances of hesitation clusters does not claim to be a tie-breaker for this discussion. However, our results point to a clear direction, comparable to those by Clark and Fox Tree (2002), differing from those by O'Connell and Kowal (2005). Silence duration does vary systematically depending on filler type, lending support to the very notion that there are indeed two distinct types of fillers.

Inter-speaker variability

Speakers produced on average 38 fillers, but as observed in an earlier study conducted on the same data by Kosmala and Morgenstern (2019), there is great inter-speaker variability in filler usage, a phenomenon which has often been observed in disfluency research. While every speaker has their own preference with regard to hesitation patterns, there seems to be no systematic influence on the variables tested. For this study it is sufficient to observe that the general tendencies observed in earlier studies hold true. But if these findings were put to practical application, such as speech synthesis, a model of one particular speaker might lead to totally different results than a model constructed on the basis of mean values from a pool of speakers.

Speaking style influence

The results may additionally be strongly influenced by the communication settings and the speaking style. Clark and Fox Tree (2002) used spontaneous face-to-face conversations, whereas O'Connell and Kowal (2005) used speech data from a trained and educated speaker being interviewed by media experts. Our analyses were conducted on interview-style data, but spontaneous nonetheless, which might be a reason for the closeness to Clark and Fox Tree's results.

Standard maximum silence

The analyses on distance revealed some further insights. Most silences, when associated with fillers, occurred in direct vicinity to them. There was a significant difference in silence duration preceding and following fillers, which requires some future attention. If this is robust, it could lend support to the notion of Standard Maximum Silence (Jefferson, 1989): there is an upper threshold for silences in conversation, and when it is exceeded, either speaker will contribute to bridge the silence. This follows the initial notion dating back from the beginning of disfluency research in the 1950s that being silent for too long puts the speaker at risk to lose the conversational floor (Maclay & Osgood, 1959). It would make sense then, to have shorter silences before fillers than after, because once the speaker has produced the filler, the listener has already been provided with a cue that conversation might continue, so there is more pressure to fill the silence when no filler has occurred yet.

Application and outlook

One practical application of the results obtained here is the extension of the hesitation insertion model for speech synthesis, which has been prototypically tested in Betz et al. (2018), which did not yet take into account the structural interplay of silences and fillers. Furthermore, the hesitation model by Betz et al. (2018) is centered on lengthening, which provides an elegant entry point for a synthetic hesitation interval, and reflects human speech production by making use of the pre-planned, but not-yet-uttered words in the articulatory buffer (Levelt, 1989). This approach receives support by the confirmed notion of longer silences after um-type fillers: the presence of a nasal sound makes this type of filler a better candidate to smoothly initiate a hesitation interval by lengthening compared to the uh-type fillers (for hesitation lengthening distribution over phone types, cf. Betz, Wagner & Voße, 2016).

For future work on these matters, it is desirable to extend the analyses started here to a dataset with phonemic annotation, so that lengthening can be

included as a third hesitation marker which might frequently cluster with silences and fillers. The hypothesis would be that hesitation lengthening clustered with fillers would presumably be associated with um-type fillers which denote a longer delay. Betz and Wagner (2016) observed that phones preceding fillers undergo the same lengthening processes as phones preceding intonation phrase boundaries. This has been explained by the presence of fillers introducing an additional intonation phrase boundary at positions not predicted by syntax, which causes the typical phrase-final lengthening to occur. However, in that study, fillers were not distinguished into uh and um types.

Conclusion

Hesitation markers occur in speech both in standalone form and in clusters. While clusters are comparably rare, it is still desirable to be able to model them adequately. In Betz, Wagner and Schlangen (2015), it was found that the more hesitation markers were included in the same synthetic utterance, the worse user ratings got. This might well be due to the fact that there are certain syntactic rules which govern how hesitation markers have to be combined. This study was a first step, investigating the much-discussed interplay of silences and fillers, for future work it is desirable to extend these analyses to include the third prototypical hesitation marker, lengthening, in order to get a full picture of the mechanisms behind hesitation clusters.

References

Benus, S., F. Enos, J. B. Hirschberg & E. E. Shriberg. 2006. Pauses in Deceptive Speech. In: R. Hoffmann & H. Mixdorff (eds.): *Proceedings of International Conference on Speech Prosody*, 2–5 May 2006, Dresden, Germany, paper number 212.

Betz, S., B. Carlmeyer, P. Wagner & B. Wrede. 2018. Interactive Hesitation Synthesis: Modelling and Evaluation. *Multimodal Technologies and Interaction* 2(1): 9. https://doi.org/10.3390/mti2010009

Betz, S. & S. López Gambino. 2016. Are We All Disfluent in Our Own Special Way and Should Dialogue Systems Also Be? In: O. Jokisch (ed.), *Studientexte zur Sprachkommunikation 81*. Dresden: TUD Press, 168–174.

Betz, S., & P. Wagner. 2016. Disfluent Lengthening in Spontaneous Speech. In: O. Jokisch (ed.), *Studientexte zur Sprachkommunikation 81*. Dresden: TUD Press, 135–144.

Betz, S., P. Wagner & D. Schlangen. 2015. Micro-Structure of Disfluencies: Basics for Conversational Speech Synthesis. In: *Proceedings of INTERSPEECH*, Dresden, Germany, 2222–2226.

Betz, S., P. Wagner & J. Voße. 2016. Deriving a Strategy for Synthesizing Lengthening Disfluencies Based on Spontaneous Conversational Speech Data. In: C. Draxler & F. Kleber (eds.), *Tagungsband Der 12. Tagung Phonetik Und Phonologie Im Deutschsprachigen Raum*, Munich, Germany, 19–22.

Clark, H. H. & J. E. Fox Tree. 2002. Using Uh and Um in Spontaneous Speaking. *Cognition* 84(1): 73–111. https://doi.org/10.1016/S0010-0277(02)00017-3

Eklund, R. 2004. *Disfluency in Swedish human-human and human-machine travel booking dialogues*. Ph.D. dissertation, Linköping University.

Finlayson, I. R. & M. Corley. 2012. Disfluency in Dialogue: An Intentional Signal from the Speaker? *Psychonomic Bulletin & Review* 19(5): 921–28. https://doi.org/10.3758/s13423-012-0279-x

Grosjean, F., & A. Deschamps. 1972. Analyse Des Variables Temporelles Du Français Spontané. *Phonetica* 26(3): 129–56. https://doi.org/10.1159/000259407

Jefferson, G. 1989. Preliminary Notes on a Possible Metric Which Provides for a 'Standard Maximum' Silences of Approximately One Second in Conversation. In: D. Roger & P. Bull (eds.), *Conversation: An Interdisciplinary Perspective*, Clevedon: Multinlingual Matters, 166–96.

Kosmala, L. & A. Morgenstern. 2017. A Preliminary Study of Hesitation Phenomena in L1 and L2 Productions: A Multimodal Approach. In: R. Eklund & R. Rose (eds.): *Proceedings of the 8th Workshop on Disfluency in Spontaneous Speech*, 18–19 August, 2017, Stockholm, Sweden, 37–40.

Kosmala, L. & A. Morgenstern. 2019. Should 'uh' and 'um' Be Categorized as Markers of Disfluency? The Use of Fillers in a Challenging Conversational Context. In: L. Degand, G. Gilquin, L. Meurant & A. C. Simon (eds.): *Fluency and Disfluency across Languages and Language Varieties*, Louvain-la-Neuve: Presses Universitaires de Louvain, 67–90.

Levelt, W. J. 1989. *Speaking*. Cambridge, MA: MIT Press.

Lickley, R. J. 2015. Fluency and Disfluency. In: A. M. Redford (ed.): *The Handbook of Speech Production*, Hoboken, NJ: Wiley Blackwell, 445. https://doi.org/10.1002/9781118584156.ch20

Maclay, H. & C. E. Osgood. 1959. Hesitation Phenomena in Spontaneous English Speech. *Word* 15(1): 19–44. https://doi.org/10.1080/00437956.1959.11659682

O'Connell, D. C. & S. Kowal. 2005. Uh and Um Revisited: Are They Interjections for Signaling Delay? *Journal of Psycholinguistic Research* 34(6), 555–76. https://doi.org/10.1007/s10936-005-9164-3

Schegloff, E. A. 2010. Some Other 'Uh(m)'s. *Discourse Processes* 47(2): 130–174. https://doi.org/10.1080/01638530903223380

Shriberg, E. E. 1994. *Preliminaries to a Theory of Speech Disfluencies*. Ph.D. dissertation, University of California at Berkeley.

Smith, V. L. & H. H. Clark. 1993. On the Course of Answering Questions. *Journal of Memory and Language* 32(1): 25–38 https://doi.org/10.1006/jmla.1993.1002.