THE SYNTACTIC, SEMANTIC, TOPIC AND SOCIOECONOMIC EFFECTS ON SILENT PAUSE DISTRIBUTION

Hong Zhang and Mark Liberman

University of Pennsylvania
zhangho@sas.upenn.edu

ABSTRACT

Silent and filled pauses in spontaneous speech have been extensively studied to test hypotheses about language production. Here, we report results from a corpus study of large collections of spontaneous telephone conversations. We specifically examined both durational and distributional properties of silent pauses as functions of syntactic structure, sentence length, semantic context, topic and socioeconomic properties such as age, gender and years of education.

Our results partially support previous hypotheses and empirical findings on this multivariate problem, while offer further details to the interplay among the multidimensional variables. It is hoped that our findings can serve as an empirical basis for future theoretical and probabilistic modeling of the structure of conversational speech in both normal and clinical populations.

Keywords: Silent pauses, duration, syntactic semantic topic and socioeconomic effects

1. INTRODUCTION

As an integral part of human speech, silent and filled pauses offer rich information about speech structuring and language production [21, 18, 7]. Efforts have been made to understand both where and how pauses occur, and what are the consequences of pauses (for example, [14, 17, 28, 25]). The volume and diversity of topics covered in silent pause research suggest that better knowledge of the multivariate nature of pauses could be essential not only in the study of the cognitive mechanism underneath speech production, but also in applications such as dialog systems and the assessment of clinical speech.

In this paper, we aim to provide a comprehensive overview of the variables that have been suggested to correlate with silent pause production using a dataset constructed from large corpora of telephone conversations. We cover the aspects that have been reported or hypothesized to have an effect on pause duration and distribution. These aspects include the syntax of utterances [10, 26], semantic contexts of pauses [3, 23, 2], discourse factor [20, 22], and socioeconomic variables of the speaker [24]. Previous studies often focused on the role played by single factors in pause production, which limits the ability for research findings to generalize beyond experiment settings. We hope the patterns explored in this paper can serve as a foundation for future theoretical and empirical research in questions related to pauses.

2. DATA

We look at two corpora of telephone conversations of English: the Switchboard [12] and Fisher corpus [6]. A random sample of 640 conversations have been selected from Switchboard (about 80 hours of speech), while a stratified sample by conversation topics have been extracted from Fisher. The Fisher sample contains one tenth of the full collection, corresponding to 1119 two-person conversations from 2069 speakers. The total amount of speech in this sample is about 180.5 hours with about 3 million words.

We select two different sources in this study mainly because the two corpora provide information suitable for answering specific questions, and the nature of the speech is comparable. For example, Switchboard has rich and accurate annotation for POS tags and syntactic category, which comes handy for answering questions related to the syntactic structure of utterances. However, clear indicator for turn segmentation is lacking in this corpus, while can be relatively easily reckoned from Fisher.

Silent pause is defined as within-turn pauses in telephone conversations. Turns are identified through the time stamps and side labels provided in corpus transcriptions. Back channel talking is eliminated by removing short filler words and segments that are shorter than four words. Using turn as the smallest speech segmentation unit, rather than utterance, avoids the subjectivity of utterance identification. Silent pause identification is based on alignment results. An HTK based forced aligner is used

to align the sample from Fisher. The performance of this aligner has been reported to be over 97% [27]. Pauses in Switchboard are found directly from the time stamps provided in the corpus. 150 ms has been used as the threshold to define silent pauses, which is in between the threshold proposed by [8] and [13]. An examination of a small (100-turn) sample shows that this threshold is able to capture the pausing phenomena while excluding silent intervals related to other linguistic processes.

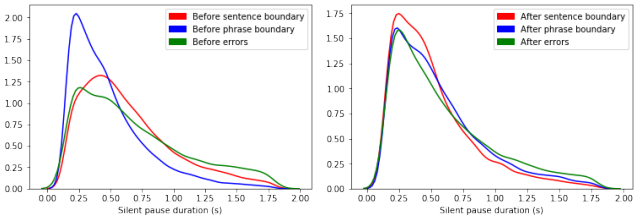
3. THE SYNTACTIC EFFECTS

In this section, we report results on the syntactic effects on silent pause frequency and duration distribution. We focus on three specific questions: What is the effect of syntactic boundary on silent pause distribution, what is the effect of turn length on silent pause duration, and what is the effect of pause location in a turn on pause duration.

3.1. Boundary effect

It has been generally agreed that the syntactic structure of utterances plays a role in both pause duration and likelihood of observing a pause [10, 17, 29]. Silent pauses are also more likely to occur adjacent to other types of disfluencies such as filled pause [23]. Therefore, we compare both the duration and frequency distribution of silent pauses in three contexts: the boundary of a tensed phrase (TP, which is referred to as *sentence*), the boundary of XP (such NP, VP or PP, referred to as *phrase*) and the boundary of other disfluent segments (referred to as *errors*). The results are based on analysis of the sample from Switchboard.

Figure 1: Silent pause duration at different phrase boundaries



(a) Before boundaries **(b)** After boundaries

Figure 1 shows the density estimation for pause duration in the three conditions. It can be noticed that the distribution is very skewed, approximating a Γ distribution rather than normal. The center of the distribution is also shifted towards the right for pauses before *sentence* and *errors* when compared to *phrase*. This suggests that pauses are longer before larger syntactic units and disfluencies than smaller constituents. A similar trend has been reported in [26], where silent pause duration is longer before longer subsequent clause and higher boundary strength in Japanese. However, this trend is not observed among pauses after the boundaries.

Figure 2: Frequency distribution of silent pauses at different phrase boundaries

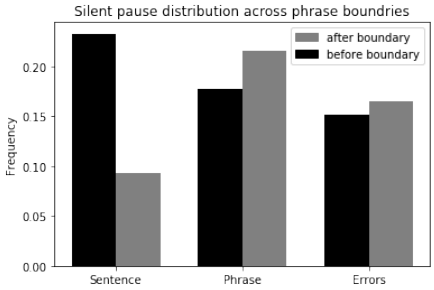


Figure 2 shows the frequency of silent pauses occurring near the boundary of the aforementioned conditions relative to the total number of pauses in the corpus. Silent pauses occur more often before a *sentence*, compared to *phrase* and *errors*. However, the relative frequency of silent pauses after *phrase* and *errors* are higher than the frequency after *sentence*, as well as the frequency before *phrase* and *errors*. This pattern suggests that pauses at different phrase boundaries may have different discourse functions, or may reflect planning problems before constituents.

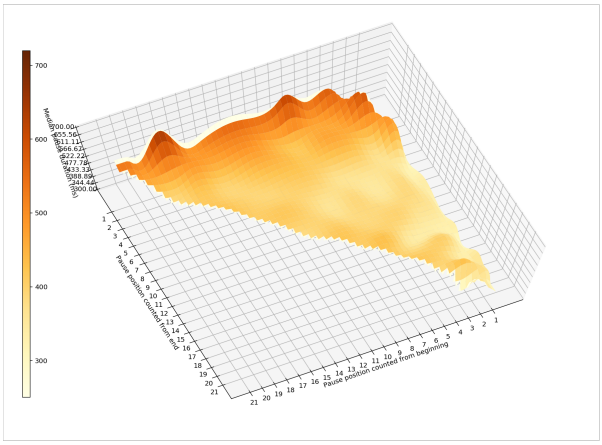
3.2. Length and turn-internal localization

In this section, we address the question of what is the three-way interaction among turn length, silent pause location and pause duration, where pause duration is the median duration of all pauses at the same position of turns of same length. The stratified sample from Fisher is used due to its convenience for and high accuracy in reconstructing turns in conversations. Pauses following fillers are excluded.

In figure 3 below, the two axes on the horizontal surface indicate the location of pauses in a turn counted from the beginning (the front axis) and end. The diagonal lines in the surface square indicate the length of the turn, all measured by the number of words. Therefore, the longest diagonal (i.e., the front-most boundary of the surface) represents the median pause duration at all possible positions in turns longer than 5 words and capped at 20 words.

Two general trends can be observed from figure 3. First, pauses in very short turns are longer, especially for turns that are only 5 to 6 words long. Second, long pauses are primarily located toward the

Figure 3: The surface plot of median silent pause duration, turn length and pause position in a turn (measured by the number of words)



end of a turn. This trend can be noticed from the uphill of the surface from the front-right corner towards the back. In addition to the effects of syntactic phrasing as discussed above, figure 3 suggests the existence of other discourse structuring functions of silent pauses in spontaneous conversations.

4. SEMANTIC EFFECT

The semantic information has been thought to reflect the lexical access process during speech production. Previous research generally suggests a relation between higher frequency and longer duration of pauses in semantic contexts that are more complex, rare and unexpected [2, 23]. With this relation in mind, here we ask if the semantic context can be useful in distinguishing different pauses.

We categorically define four types of silent pauses based on the overall distribution of pause duration in our corpora, with reference to [29, 5, 4]'s work on overall pause duration distribution. The thresholds for four types silent pauses are shown in table 1. A 2000ms upper bound is applied to exclude potential non-turn-internal pauses and other pausing phenomena. The results are based on joint Switchboard and Fisher sample.

Table 1: Definition of types of silent pauses

pause category	criteria (ms)
short pause	$150 \leq$ pause < 400
mid pause	$400 \leq$ pause < 600
mid-long pause	$600 \leq$ pause < 800
long pause	$800 \leq$ pause < 2000

To uncover the latent semantic dimensions that can capture the lexical context for pauses, we treat each "pause label" as words, and apply Latent Semantic Analysis (LSA), implemented using word2vec [19]. The full in-sample transcripts are used for training, as filtering out stop words may lose too much information that might be critical for conversational speech. The model is trained with a window of ± 5 words, and word-word vectors are reduced to 100 dimension dense vectors. Metrical Multidimensional Scaling (MDS) is used to project clusters in word-vector space onto a 2D plane. Two variants of filled pause, "uh" and "um", are also included in this pause analysis as reference points.

Figure 4: Word embedding for different pauses projected in 2D

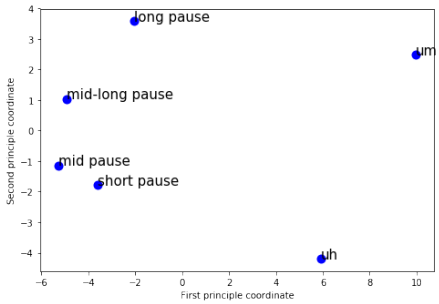


Figure 4 shows the clustering result. A clear separation between silent and filled pauses can be observed along the first principle coordinate. The second principle coordinate appears to track the continuous change from short to long pauses, with clear gaps between "long pause" and other pause categories. A secondary separation between mid-long pause and mid and short pauses can also be argued. Interestingly, in the second principle coordinate, the separation between "um" and "uh" parallels with the separations along pause duration. It can be argued that the filled pause "um" shares certain similarities with longer silent pauses in terms of its semantic contexts. Similar argument can be made for "uh" and shorter pauses. However, the semantic correlates for silent and filled pauses are still rather different. Our follow-up work will look at the exact semantic contexts for different pause categories in more detail.

5. TOPIC AND SOCIOECONOMIC EFFECTS

Discourse and speakers' idiosyncratic features have also been treated as influencing factors in disfluencies. It has been reported that academic field could affect the fluency of college lectures [22], and intelligence score correlates with the fluency of verbal communication [9]. More attention has also been paid to filled pause distribution, such as [1, 11] Here,

we report our regression analyses on the effects of conversation topic, speaker age, gender and years of education on the rate and duration of silent pauses using the Fisher sample.

5.1. Model specification

We use median regression to probe the relation between pause rate and duration, and topic and speaker features. Median regression is considered more appropriate due to the highly skewed distribution of pause duration and frequency.

Two models are explored, with per-speaker *median silent pause frequency*, measured as the number of silent pauses per turn, and per-speaker *median silent pause duration*, measured in millisecond, as response variables respectively. Explanatory feature space has been constructed to include predictors that can as exhaustively explain the variance in response variables as possible to minimize the potential endogeneity caused by unobserved effects. Thus, we consider the following set of predictors as shown in table 2.

The total number of observations for regression analyses is 2238. Since in cases where a same speaker participated in several conversations, the topics are distinct, repeated observations of same speakers are not collapsed.

Table 2: Explanatory variable list in regressions

predictor	value
$utterance\ length$	average no. of words per turn
$topic$	assigned topic in conversation
$dialect$	speaker's dialect background
$education$	self-reported yrs. of education
age	speaker age in conversation
$gender$	gender identified by voice
$sex \times accent$	accent: American or not

5.2. Feature generation

The original 40 topics are collapsed to 17 based on the cosine similarity between pairs of conversation contents. Dialects are inferred from the state of origin of speakers and regrouped to seven major American English dialects, plus Canadian and foreign accents.

5.3. Regression results

The median regressions were performed using R's *quantreg* package [16]. Standard errors and p-values were calculated with the Sandwich formula[15] to account for heteroskedasticity.

5.3.1. Effects on silent pause rate

Among the predictors of interests, only *topic* shows significant effect on the rate of silent pause ($p = .000, F = 4.733, DF = 16, 2171$). This result suggests that certain conversation topics are likely to induce more pauses in the dialog. However, the rate of silent pauses in conversations is not dependent upon speaker-specific features.

5.3.2. Effects on silent pause duration

In this model, *topic* ($p = .0003, F = 2.707, DF = 16, 2171$), *education* ($p = .0055, se = 0.636, \beta = -1.766$) and $sex \times accent$ ($p = .000, F = 15.405, DF = 3, 2171$) show significant effects on median silent pause duration. $dialect$ is marginally significant ($p = .062, F = 1.926, DF = 7, 2171$) if .05 threshold is chosen. Together with the previous model, we see that the broader contexts of conversations, which is proxied by *topic* in the model, affect both the frequency and duration of silent pauses. In addition, speaker's accent and its interaction with gender also have effects on pause duration, but not frequency. Speaker's education level is also suggested to affect pause duration.

The two regression models confirm the hypothesis that broader conversational context could impact silent pause production, in terms of both pause frequency and duration. On the other hand, certain speaker-specific features only affect pause duration but not frequency. The exact interactions among these variables warrant further investigation.

6. CONCLUSION

In this paper, we mainly reported results from a preliminary exploration of linguistic and socioeconomic effects on silent pause duration and frequency distribution. This exploratory corpus analysis provides further details to the existing knowledge about the relationship between pause duration, speech length, and syntactic structure. We also demonstrated that the semantic contexts can be informative about the length distinction among silent pauses. A parallel between silent and filled pauses can also be drawn in this manner. Broader discourse factors, such as topic, and speaker features may also play a role in the making of silent pauses.

7. REFERENCES

[1] Acton, E. K. 2011. On gender differences in the distribution of um and uh. *University of Pennsylvania Working Papers in Linguistics* 17(2), 2.

[2] Arnold, J. E., Kam, C. L. H., Tanenhaus, M. K. 2007. If you say thee uh you are describing some-

thing hard: the on-line attribution of disfluency during reference comprehension. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 33(5), 914.

[3] Beattie, G. W., Butterworth, B. L. 1979. Contextual probability and word frequency as determinants of pauses and errors in spontaneous speech. *Language and Speech* 22(3), 201–211.

[4] Campione, E., Véronis, J. 2002. A large-scale multilingual study of silent pause duration. *Speech Prosody 2002, International Conference*.

[5] Campione, E., Véronis, J. 2005. Pauses and hesitations in French spontaneous speech. *Disfluency in Spontaneous Speech*.

[6] Cieri, C., Miller, D., Walker, K. 2004. The Fisher Corpus: a Resource for the Next Generations of Speech-to-Text. *LREC* volume 4 69–71.

[7] Clark, H. H., Tree, J. E. F. 2002. Using uh and um in spontaneous speaking. *Cognition* 84(1), 73–111.

[8] Eklund, R. 2004. *Disfluency in Swedish human–human and human–machine travel booking dialogues*. PhD thesis Linköping University Electronic Press.

[9] Engelhardt, P. E., Nigg, J. T., Ferreira, F. 2013. Is the fluency of language outputs related to individual differences in intelligence and executive function? *Acta psychologica* 144(2), 424–432.

[10] Ferreira, F. 1991. Effects of length and syntactic complexity on initiation times for prepared utterances. *Journal of Memory and Language* 30(2), 210–233.

[11] Fruehwald, J. 2016. Filled pause choice as a sociolinguistic variable. *University of Pennsylvania Working Papers in Linguistics* 22(2), 6.

[12] Godfrey, J. J., Holliman, E. C., McDaniel, J. 1992. SWITCHBOARD: Telephone speech corpus for research and development. *IEEE International Conference on Acoustics, Speech, and Signal Processing, 1992. ICASSP-92., 1992* volume 1. IEEE 517–520.

[13] Goldman-Eisler, F. 1958. Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology* 10(2), 96–106.

[14] Holmes, V. M. 1988. Hesitations and sentence planning. *Language and Cognitive Processes* 3(4), 323–361.

[15] Huber, P. J., others, 1967. The behavior of maximum likelihood estimates under nonstandard conditions. *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* volume 1. University of California Press 221–233.

[16] Koenker, R. 2012. quantreg: Quantile Regression. R package version 4.79.

[17] Krivokapić, J. 2007. Prosodic planning: Effects of phrasal length and complexity on pause duration. *Journal of Phonetics* 35(2), 162–179.

[18] Levelt, W. J. 1993. *Speaking: From intention to articulation* volume 1. MIT press.

[19] Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., Dean, J. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems* 3111–3119.

[20] Moniz, H., Batista, F., Mata, A. I., Trancoso, I. 2014. Speaking style effects in the production of disfluencies. *Speech Communication* 65, 20–35.

[21] Rochester, S. R. 1973. The significance of pauses in spontaneous speech. *Journal of Psycholinguistic Research* 2(1), 51–81.

[22] Schachter, S., Christenfeld, N., Ravina, B., Bilous, F. 1991. Speech disfluency and the structure of knowledge. *Journal of Personality and Social Psychology* 60(3), 362.

[23] Shriberg, E., Stolcke, A. 1996. Word predictability after hesitations: a corpus-based study. *ICSLP 96., Fourth International Conference on Spoken Language Processing.* volume 3. IEEE 1868–1871.

[24] Tottie, G. 2011. Uh and um as sociolinguistic markers in British English. *International Journal of Corpus Linguistics* 16(2), 173–197.

[25] Tsiamtsiouris, J., Cairns, H. S. 2013. Effects of sentence-structure complexity on speech initiation time and disfluency. *Journal of Fluency Disorders* 38(1), 30–44.

[26] Watanabe, M., Kashiwagi, Y., Maekawa, K. 2015. The relationship between preceding clause type, subsequent clause length and duration of silent and filled pauses at clause boundaries in japanese monologues. *The 7th Workshop on Disfluency in Spontaneous Speech (DiSS 2015)*.

[27] Yuan, J., Liberman, M. 2008. Speaker identification on the SCOTUS corpus. *Journal of the Acoustical Society of America* 123(5), 3878.

[28] Zellner, B. 1994. Pauses and the temporal structure of speech. In: *Zellner, B.(1994). Pauses and the temporal structure of speech, in E. Keller (Ed.) Fundamentals of speech synthesis and speech recognition.(pp. 41-62). Chichester: John Wiley.* John Wiley 41–62.

[29] Zvonik, E., Cummins, F. 2003. The effect of surrounding phrase lengths on pause duration. *Eighth European Conference on Speech Communication and Technology*.