INTERSPEECH 2009 BRIGHTON

Variability and stability in collaborative dialogues: turn-taking and filled pauses

Štefan Beňuš

Constantine the Philosopher University, Nitra, Slovakia and Slovak Academy of Sciences, Bratislava, Slovakia

sbenus@ukf.sk

Abstract

Filled pauses have important and varied functions in turn-taking behavior, and better understanding of this relationship opens new ways for improving the quality and naturalness of dialogue systems. We use a corpus of collaborative task oriented dialogues to provide new insights into the relationship between filled pauses and turn-taking based on temporal and acoustic features. We then explore which of these patterns are stable and robust across speakers, which are prone to entrainment based on conversational partners, and which are variable and noisy. Our findings suggest that intensity is the least stable feature followed by pitch-related features, and temporal features relating filled pauses to chunking and turn-taking are the most stable.

Index Terms: filled pauses, turn-taking, speaker variation

1. Introduction

The use of filled pauses, turn-taking behavior, and their interaction offer an important window into the cognitive dynamic system of human conversations. Improved understanding of these systems brings great potential for improving the quality and naturalness of dialogue systems and interactive voice response applications.

Filled pauses (FPs) such *um* or *uh* signal multiple communicative functions in both the production and perception of speech; see [1] for a recent review. In production, speakers tend to use FPs systematically to mark structural junctures at prosodic, syntactic, pragmatic, or discourse levels; and to signal cognitive load and/or planning difficulties associated with a choice. Listeners were also shown to be highly sensitive to the occurrence of FPs in speech. FPs facilitate the process of comprehension by helping listeners better predict information in upcoming speech, and by helping memory retention of words preceded by FPs.

Turn-taking behavior is a complex dynamic cognitive system that determines who speaks when. Interlocutors constantly produce and monitor turn-yielding and turn-holding cues for a potential entry into the conversation. Many of these cues are assumed to be dynamically accommodated or entrained among the conversational partners, e.g. [2]

Interaction between FP use and turn-taking is also communicatively meaningful. Turn-initial FPs belong among 'entry devices' that facilitate both production and perception of linguistic material, because they allow speakers to think about and plan the intended message, and they let listeners get ready to perceive important content [3]. FPs are also used to actively assume the floor in conversations, and mark speakers' intentions to hold the floor in dialogues [4], and differ phonetically based on their floor-management functions and turn-positions [5, 6].

Finally, [7] showed that FPs are more frequent in spontaneous speech than in one-way speech with mechanical control of turns, which is similar to current dialogue system applications. Therefore, [7] concluded, FPs are necessary in managing spontaneous-like conversations. FPs are also effective in real turn-taking applications. For example, a turn-initial *conversational filler* by a robot significantly improved the user's impression of longer response times [8].

In this paper, we aim at better understanding of the relationship between FPs and turn-taking in human interactions by pursuing two goals. First, we use a corpus of collaborative task oriented dialogues to provide new insights on the relationship between FPs and turn-taking based on the quantitative analysis of real spontaneous speech. Second, the design of our corpus allows us to explore stable and variable features in this domain for further use and applications. We ask which of the patterns are stable and robust across speakers, which are prone to entrainment based on conversational partners, and which are variable/noisy. In this way, stable behavior across speakers is useful for general applicability. Stable features within a conversation but different across interlocutors are potentially useful for 'customizing' and user accommodation. And, knowing which features are variable even in a single conversation, and thus not useful for functional goals, might be useful for increasing naturalness of synthesized speech in dialogue systems.

2. The corpus

The relationship between turn-taking and the use of FPs in this paper is analyzed with the data from the Columbia Games Corpus [9, 10]. The corpus consists of 12 dyadic spontaneous task-oriented conversations elicited from 13 speakers of standard American English (7 males and 6 females). Subjects used separate laptops to play two types of collaborative games (CARDS and OBJECTS). 11 subjects played with two different partners in two different sessions, and 2 played a single session. The dialogues were recorded in a soundproof booth. Subjects could not see each other due to a curtain, which effectively limited all non-verbal interactions.

In three variations of the CARDS games, subjects received points for finding cards depicting the same objects on their laptop screens. Players took turns and one described a card on her screen, while the other searched for a full or partial match on his board. In the OBJECTS games, one player described the position of a target object with respect to other fixed objects on her screen, while the other tried to move his representation of the target object to the same position on his own screen. Points were given based on the proximity of the target object to its correct location. Both games were designed to encourage discussion, and subjects were motivated to exchange as much information as possible by promising them money for accumulated points. The subjects switched the roles of Describer and Searcher/Placer repeatedly.

All interactions were recorded, digitized, and downsampled to 16K. The recordings were orthographically transcribed, and words were aligned to the source acoustic

signal by hand. There is the total of 70,259 words (tokens) and 2037 unique words (types) in the corpus in over 9 hours of speech. The transcripts yielded 791 *ums*, 861 *uhs*, and 141 *mms*, for a total of 1793 tokens treated as FPs. The rate of FP use in this corpus is thus 2.5%, which is comparable to similar corpora [11].

Chunks of speech were determined automatically as pause-defined units within a single turn with the duration of pause at least 50ms. About half of the corpus (all of the OBJECTS and 60min of CARDS games) was intonationally transcribed using the ToBI conventions [12]. All continuous acoustic features for pitch, intensity, and formant values were automatically extracted from the signal using Praat [13].

3. Results

3.1. Filled pauses, turns, and chunks

3.1.1. Descriptive observations

Peripheral positions – when FPs are flanked by a silent pause from one or both sides – are dominant. Table 1 shows that only 16.7% of all FPs are not preceded or followed by some silence; these are turn-medial chunk-medial FPs in the last row. Additionally, a third of all FPs (32.9%) start a turn. These observations point to an important delimitative function of FPs that is closely linked to turn-taking because these peripheral positions suggest several floor-management functions.

Turn-initial positions suggests that these FPs initiated successful floor-grabbing or that the FP has a pre-start function that allows the speaker some time for planning and the listener for tuning in. Turn-final positions are the least frequent (3.3%); hence, the turn-yielding function of FPs is not supported in our corpus. Turn-only FPs are rare but suggest unsuccessful floor-grabbing, or perhaps, floor-yielding hesitations that could serve as prompts for more input from the interlocutor.

Chunk-final (16.7%) and chunk-only (27%) in turn-initial or medial positions are very frequent and suggest the floor-holding function. This is because these FPs are followed by silent pauses during which the other interlocutor did not assume the floor. Finally, the chunk-initial position of FPs within a turn (16%) suggests a plain hesitation pause.

Table 1. *Distribution of filled pauses within turns(T) and chunks(Ch); ini(tial), fin(al), med(ial).*

Turn/Chunk	Um	uh	mm	Total
T-only/Ch-only	24	11	38	73
T-ini/Ch-only	130	44	21	195
T-ini/Ch-ini	98	177	47	322
T-fin/Ch-only	20	3	3	26
T-fin/Ch-fin	16	17	0	33
T-med/Ch-only	187	61	9	257
T-med/Ch-ini	82	187	18	287
T-med/Ch-fin	171	127	2	300
T-med/Ch-med	63	234	3	300
Total	791	861	141	1793

The table also shows that *mm* is most common turn-initially, in which it may be ambiguous between FP and acknowledgement functions. The main difference between *um* and *uh* is that *um* is consistently more likely to be followed by a silent pause than *uh*, which is the difference between chunk-only and chunk-initial positions.

3.1.1. Speaker variation

Since constructing similar tables for each speaker based on his/her interlocutor would not be informative, we only looked at turn-initial FP positions given their communicative salience. Given the 1/3 vs. 2/3 split for turn-initial FPs, 1 speaker (103) had a significant reversal (67% of turn-initial FPs), 4 speakers had roughly even split, 1 had significantly lower ratio (11% of turn-initial FPs), and the remaining 7 speakers followed the general pattern. Testing the effect of interlocutor, only 2 of 11 speakers significantly changed the ratio of turn-initial FPs in the two sessions: for speaker 103, 50% of FPs were turn-initial with interlocutor 104 but 78% with 111; and for speaker 112, 27% FPs were turn-initial with interlocutor 110 but 66% with 113. Hence, the inclination of speakers to start a turn with an FP is reasonably stable across speakers and even more stable within speakers communicating with different interlocutors.

3.2. Filled pauses and temporal features

3.2.1. Descriptive observations

[14] argued that *um* and *uh* are different lexical items in part based on the function that they have. *Um* signals deeper planning problems while *uh* tends to signal lexical retrieval problems. This suggestion was corroborated in [14], and also in [6] in a corpus of interviews, where they found that *ums* were more likely followed by longer silent pauses than *uhs*. However, it is not clear if these differences are due to the length of the FPs or the lexical difference between them.

First, we tested if there is a positive correlation between the lengths of FPs and the following silent pauses. We found a significant positive correlation overall, with the strongest effect in the turn-initial position, $r(195) = 0.38$, $p < 0.001$. But, the lexical difference in FPs does not seem to play a role, since the correlation for all FPs followed by a silent pause other than in the final position was much stronger for *uh* than for *um+mm*, $r(232) = 0.43$, $p < 0.001$ and $r(520) = 0.19$, $p < 0.001$ respectively. Additionally, both nasal FPs in turn-initial position, and non-turn-initial *uhs*, show roughly similar correlations between the length of FP and the following silent pause, $r(151) = 0.41$, $p < 0.001$ and $r(188) = 0.48$, $p < 0.001$ respectively. Hence, although *ums* are more likely to be followed by a silent pause than *uhs*, as discussed with Table 1, once an FP is followed by silence, the positive correlation between their durations holds irrespective of the 'lexical' difference.

We also tested for the correlation between FPs and silent pauses that precede them in turn-internal positions. Here, the effect was less robust, yet still significant, $r(570) = 0.15$, $p < 0.001$. Testing for separate effects of FP type, the correlation for oral *uh* increased while that for nasal FPs decreased and became non-significant, $r(251) = 0.29$, $p < 0.001$ and $r(319) = 0.05$, $p = 0.48$ respectively.

3.2.2. Speaker variation

Given the significant correlations reported above, we calculated the 'following pause ratio' – duration of the FP divided by the following silent pause. On average, FPs were 1.7 times longer than the following silent pause. Although between-speaker effect was significant, the post-hoc tests showed that this was due to only one subject (109) whose

values were significantly greater than the rest. In turn-initial FPs, the significant between-speaker effect disappeared. Finally, testing for within-speaker differences based on the interlocutor, only 1 out of 11 subjects showed a significant effect. Hence the metrical feature describing the pause ratio is rather stable, especially in the salient turn-initial position.

Similarly to the following pause, we also calculated 'preceding pause ratio'. On average, FPs were 2 times longer than the preceding silent pause. Neither between-speaker effects, nor the effects of conversational partner on each speaker, were significant. Hence, preceding pause ratio is also a relatively stable feature describing the temporal aspects of FP use.

The means of the two pause ratios for each speaker are shown in Figure 1. Although the pattern is complex, we can observe congruence in the two values, which points to some saliency in the production of adjacent filled and silent pauses. Speaker 109 is again clearly different from the rest since she has much longer following than the preceding silences in relation to the length of FPs. The remaining speakers either have very similar ratios (105, 107, 110, 111, 113), or have significantly longer preceding than the following silences.

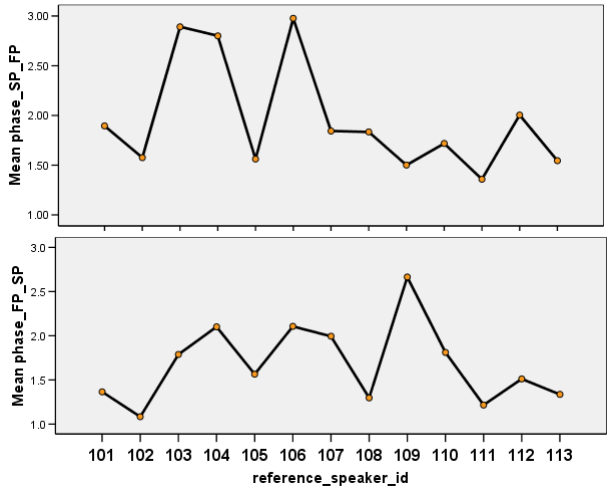

Figure 1: *Ratio of the durations for the preceding silent pause (SP) and the filled pause on the top, and following SP and filled pause on the bottom.*

3.3. Acoustic features of filled pauses

3.3.1. Continuous features and speaker variation

Recall that there were 13 speakers and 11 played the game with two different partners. We analyzed data from these 11 subjects for variance across speakers and also for variance within speakers based on the conversational partner. F-values in one-way Anova tests combined with post-hoc Tukey tests showed that normalized intensity and raw duration of FPs are the acoustic features that vary the most across speakers; $F(10, 1587) = 25.1$ for maximum intensity, and $F(10, 1615) = 15.8$ for duration. All other F-values were less than 4.

Similar results were obtained when we tested for significant differences for each speaker separately with the interlocutor as an independent variable. The results are summarized in Table 2 that shows, for example, that female speaker 103 has greater intensity, pitch, and duration when she talks with female interlocutor 111 than with male interlocutor 104. The table shows that intensity produces the greatest variation based on the interlocutor while the mean and maximum pitch, and duration are rather stable. Pitch slope features never reach significance and could be considered stable as well.

Table 2. *Influence of interlocutor (Sp-O) on duration, F0, formants and normalized intensity. '^' and 'v' show the direction of difference, '*' marks significance at $p < 0.05$, '**' at $p < 0.001$. Red speakers are female, blue ones are male.*

Sp-R	Sp-O	Dur	$F0_{max/mean}$	F1/F2	$INT_{max/mean}$
101	102 109				
102	101 105				^**
103	104 111	^*	^**(max)	^*(F1)	^**
105	102 106			v*(F1)	v*
106	105 107				^**
107	106 108	v*	^*(mean)		v*(max)
108	107 109				v*(max)
109	108 101				
110	111 112		v*(mean)	v*	
111	110 103		^*(mean)	v*(F1)	
112	110 113	^*		v*(F1)	v*(mean)

Rather surprising is the difference observed in the first formant (F1) for five speakers (all of them females). F1 typically relates to the tongue body height and is primarily responsible for the difference between /eh/, /uh/ and /ah/ vowel qualities. Although nasality in FPs, for which we don't have temporal annotations, may have confounded this finding, the reported results hold after excluding *mm* and mostly apply for formants extracted both from the temporal midpoint as well as the first quarter of the FPs. This shows that not only is the vowel quality highly variable across speakers in FPs, but it could be influenced by conversational partners.

These differences may arise from multiple factors that could potentially differ in the two conversations such as the frequency of *uh* vs. *um*, or FP positions related to chunks and turns. In fact, we already observed in section 3.1.1 that speaker 103 was one of only two speakers for which the ratio of turn-initial FPs significantly changed in the two sessions: 50% FPs were turn-initial for 104 but 78% for 111. Since turn-initial FPs tend to be longer, higher in pitch and intensity, the interlocutor effect on FP pattern with speaker 103 may be traced to turn and chunk related positions.

Finally, we looked at the relationship between pitch and intensity of the turn-initial FPs and final portions of the preceding turn of the interlocutor. [15] found that the pitch of turn-medial FPs can be predicted from the last pitch peak in the preceding chunk of the same speaker using a simple linear model. We were interested to see if the mean or maximum pitch of turn-initial FPs might be similarly related to turn-final pitch peaks from the interlocutor. In the effort to include as many tokens as possible, we used normalized automatically extracted means and maxima from the FPs and from the last 0.5s and 1s of the preceding turns. We found no significant correlation between the FP and preceding turn for pitch features but several moderate, yet significant, positive correlations for intensity features; for example maximum intensity of the FP correlated with maximum intensity in the

last 500ms of the preceding turn, r (294) = 0.26, p < 0.001. This suggests that speakers are inclined to accommodate intensity of turn-initial FPs to that of their interlocutors.

In sum, intensity is a truly variable and noisy feature of FP use prone to entrainment from interlocutors, plain FP duration and vowel quality are also variable, and other acoustic features show rather stable behavior across and within speakers.

3.3.2. Discrete features

About half of the FPs (N = 900) were ToBI transcribed, and 84% of them were judged to be prosodically prominent by being labeled with a pitch accent. We expected that turn-initial FPs would be more likely to receive a pitch accent, and more likely to have a (H)igh accent due to an expanded pitch range associated with new prosodic or discourse units. Excluding 13 cases of turn-internal FPs labeled as pitch accented but uncertain about the pitch accent type (X*), Pearson chi-square test showed that turn-position significantly affected pitch accent type, $X^2(2, N = 887) = 40.7$, $p < 0.001$. As Figure 2 shows, turn-initial FPs were significantly more likely to be produced with a H(igh) pitch accent while turn-internal FPs were more likely to have a L(ow) or no pitch accent.

Interestingly, looking at nasal (*um* + *mm*) and oral (*uh*) FPs, the association of H* with the turn-initial positions is due to the nasal FPs only, while the tendency for turn-medial FPs to have L* or no accent is due to the oral FPs. Additionally, one-way Anova tests with normalized continuous features showed that turn-initial FPs were produced significantly louder and higher than the turn-medial ones, and no difference was observed for pitch slope features.

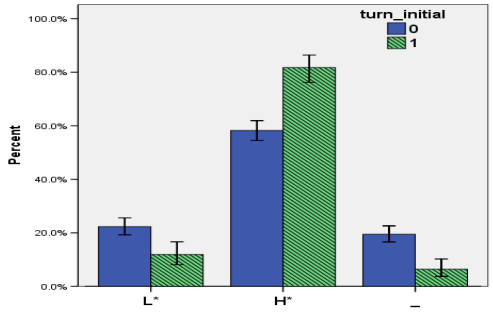

Figure 2: *Simplified pitch-accent type in turn-initial and non-initial positions.*

ToBI phrase accents and boundary tones also showed a difference between turn-positions. Turn-initial FPs were more likely to have a plateau contour than a falling one while turn-medial FPs tended to be equally split between these two most common contours.

4. Conclusions

We analyzed the relationship between FPs and turn-taking based on temporal and acoustic features, and explored which of these patterns are stable and robust across speakers, which are prone to entrainment based on conversational partner, and which are variable and noisy. Our findings suggest that intensity is the least stable feature varying both within and across speakers, and even prone to local entrainment in the vicinity of turn-exchanges. Pitch-related features displayed less speaker variability, especially in the salient turn-initial position. Although plain duration of FPs displayed significant speaker variation, the relational features describing the temporal characteristics of FPs together with surrounding silences were the most stable features, and suggest that FPs are also related to the rhythmical structure of speech.

Our findings also support some claims in the literature that nasal and oral FPs in American English are used differently in relation to turn-taking behavior. However, the notion of categorical 'lexical' difference between FPs seems to be reducible to continuous differences: *ums* are more likely to be followed by a silent pause than *uhs*, but once an FP is followed by silence, the positive correlation between their durations holds irrespective of the 'lexical' difference. It is important to keep in mind, however, that FP features of American English might not be straightforwardly extended to other Germanic languages, maybe not even into British English [16], which warrants future research in this domain.

5. Acknowledgements

The work presented in this paper was supported in part by the Slovak ministry of education grant KEGA 3/6399/08, the Slovak Research and Development Agency project APVV-0369-07, and A. W. Mellon scholarship by the IASH at the University of Edinburgh. This research was done in part in collaboration with J. Hirschberg and A. Gravano at Columbia University, USA.

6. References

[1] Corley, M. and Stewart, O. W., "Hesitation disfluencies in spontaneous speech: The meaning of um", Language and Linguistics Compass, 4: 589-602, 2008.

[2] Garrod S., and Pickering, M.J., "Why is conversation so easy?", Trends in Cognitive Sciences 8: 8-11, 2004.

[3] Sacks, H., Schegloff, E., and Jefferson, G. "A simplest systematics for the organization of turn-taking for conversation", Language, 50:696–735, 1974.

[4] Stenström, A., "Pauses in monologue and dialogue", in J. Svartvik [Ed], London-Lund Corpus of Spoken English: Description and Research, Lund University Press, 1990.

[5] Local, J. and Kelly, J., "Projection and "silences": Notes on phonetic and conversational structure", Human Studies 9: 185-204, 1986.

[6] Benus, S., Enos, F., Hirschberg, J., Shriberg, E., "Pauses and deceptive speech", in *Proceedings of 3rd International Conference on Speech Prosody*, Dresden, 2006.

[7] Taboada, M., "Spontaneous and non-spontaneous turn-taking", Journal of Pragmatics 16(2-3): 329-360, 2006.

[8] Shiwa, T., Kanda, T., Imai, M., Ishiguro, H., Hagita, N., "How quickly should communication robots respond?", HRI Proceedings, 153-160, 2008.

[9] Gravano, A., Turn-Taking and Affirmative Cue Words in Task-Oriented Dialogue. PhD thesis, Columbia University, 2009.

[10] Gravano, A., Benus, S., Hirshberg, J., Mitchell, S., Vovsha, I., "Classification of Discourse Functions of Affirmative Words in Spoken Dialogue", *Proceedings of Interspeech*, 1613-1616, 2007.

[11] Shriberg, E., "To "Errrr" is Human: Ecology and Acoustics of Speech Disfluencies. JIPA 31(1): 153-169, 2001.

[12] Beckman, M. E., Hirschberg, J., and Shattuck-Hufnagel, S., "The original ToBI system and the evolution of the ToBI framework", in S.-A. Jun, [ed] Prosodic Typology: The Phonology of Intonation and Phrasing, 9-54. OUP, 2005.

[13] Boersma, P, and Weenink, D., Praat: Doing phonetics by computer, http://www.praat.org, 2009.

[14] Smith, V. L. and Clark, H. H., "On the course of answering questions", Journal of Memory and Language 32: 25-38, 1993.

[15] Shriberg, E., and Lickley, R., "Intonation of clause-internal filled pauses", Phonetica 50: 172-179, 1993.

[16] Leeuw, E., de., "Hesitation markers in English, German, and Dutch", *Journal of Germanic Linguistics* 19(2): 85-114, 2007.