Very short utterances and timing in turn-taking

Mattias Heldner, Jens Edlund, Anna Hjalmarsson, Kornel Laskowski

KTH Speech, Music and Hearing, Stockholm, Sweden

mattias@speech.kth.se, edlund@speech.kth.se, annah@speech.kth.se, kornel@cs.cmu.edu

Abstract

This work explores the timing of very short utterances in conversations, as well as the effects of excluding intervals adjacent to such utterances from distributions of between-speaker interval durations. The results show that very short utterances are more precisely timed to the preceding utterance than longer utterances in terms of a smaller variance and a larger proportion of no-gap-no-overlaps. Excluding intervals adjacent to very short utterances furthermore results in measures of central tendency closer to zero (i.e. no-gap-no-overlaps) as well as larger variance (i.e. relatively longer gaps and overlaps).

Index Terms: Human speech production, Prosody

1. Introduction

Distributions of between-speaker-intervals in conversation – intervals of silence or *gaps*, of simultaneous speech or *overlaps*, and perfectly timed *no-gap-no-overlaps* in transitions from one speaker to another [1] – are useful for a variety of purposes ranging from basic research issues such as describing the timing of human interactive behavior to applied research issues such as deciding on suitable places to speak for spoken dialogue systems. This work explores, in particular, the timing of very short utterances (VSUs; < 1s; such as most backchannels) relative to neighboring utterances by contrasting between-speaker intervals adjacent to VSUs with those delineated by longer utterances (NONVSUS) [2]. Timing is gauged by the central tendency and variance of the between-speaker interval distributions, as well as by the relative proportions of perceived gaps, overlaps and no-gap-no-overlaps.

The distinction between VSUs and longer utterances is also used to investigate what effects come from excluding between-speaker intervals adjacent to VSUs from distributions. We and others have previously investigated between-speaker interval distributions without excluding any utterances [e.g. 3, 4], but excluding certain utterances, in particular backchannels, from studies is also a common practice [e.g. 5].

The literature does not unambiguously answer whether there is a difference in the timing of VSUs compared to longer utterances, and whether the exclusion of between-speaker intervals adjacent to VSUs affects distributions of between-speaker intervals. By many definitions, backchannels can be inserted into another's speech without interrupting [6], which may render their timing less critical and hence the variance of preceding intervals larger. On the other hand, they are typically described as short and quiet low-content utterances with a relatively homogenous function – merely indicating that the speaker who utters them is following, understanding and encouraging the other speaker to continue [6-9]. The latter characteristics may result in a lower cognitive load which might make possible a more precise timing (i.e. smaller variance and/or lower central tendency) relative to the previous or ongoing utterance compared to NonVSUs with relatively more content and variation, and an associated higher cognitive load. Furthermore, a recent study [10] observed a shift towards relatively fewer overlaps and more gaps in speaker changes involving Acknowledge Moves (which closely resembles backchannels) compared to other dialogue moves in the HCRC Map Task Corpus for Scottish English [11].

In the present paper, we extend our previous analyses of between-speaker intervals by including the operationally defined and automatically extractable distinction between VSUs and NONVSUS, which has been shown to roughly correspond to the distinction between backchannels and non-backchannels in manually annotated data [12]. We describe and evaluate a new method for speech activity detection [13] which forms the basis for the annotations of interactional phenomena. We interpret the between-speaker intervals with respect to the detection thresholds for gaps and overlaps established in [14].

2. The Spontal Corpus

The speech material used in this work was drawn from the Spontal corpus [15]. The corpus consists of recordings of audio, video, and three-dimensional motion capture from around 120 half-hour sessions of spontaneous two-party face-to-face conversations in Swedish. A subset of this corpus is split into a TRAINSET of 23 dialogues, a DEVSET of 6 dialogues, and an EVALSET of 6 dialogues. Each recording is formally divided into three consecutive 10+ minute blocks. Here, we used the close-talk microphone recordings from the first two blocks of the dialogues in the TRAINSET only, for a total of 8 hours and 13 minutes of recordings.

3. Annotation of interactional phenomena

The Spontal data was annotated for interactional phenomena using an extended version of the computational model of interaction described in [3], building on a tradition of computational models of interaction [see e.g. 13 and references mentioned therein].

3.1. Speech activity detection

First, we performed automatic speech activity detection (SAD) for both speakers simultaneously, employing the relatively low probability of the occurrence of overlap to reduce the errors caused by crosstalk [cf. 13, ch. 11]. The particular decoder used here was a hidden Markov model which had a 100 ms frame step; a 200 ms frame size; minimum speech and silence duration constraints of 200 ms; an unsupervised acoustic model of the log-energy from both channels; a supervised acoustic model of the log-energy, MFCC, and first- and second-order differences for each channel individually. The two acoustic models were interpolated linearly in the log-likelihood domain.

These characteristics were selected based on a comparison of performance against manual speech/non-speech segmentation, available for a 60 minute subset of the corpus. A miss rate and false alarm rate of 3.35% and 10.24%, respectively, were achieved. (In optimizing system parameters, we preferred low miss rates rather than low false alarm rates, all other things being equal; the segmentation was intended to aid in subsequent manual transcription of the corpus.)

The SAD produced a segmentation of each speaker state sequence into TALKSPURTS and PAUSES. TALKSPURTS were defined as a minimum of two contiguous speech frames (i.e. 200 ms, as enforced by the decoding topology) by one party that were preceded *and* followed by a minimum of two contiguous silence frames from the speaker. Similarly, PAUSES were defined as a minimum of two contiguous silence frames from that speaker.

3.2. VSUS and NONVSUS

The TALKSPURTS were subdivided into very short utterances (VSUS) and their complement (NONVSUS) based on their duration. TALKSPURTS between 2 and 10 frames in duration (i.e. 200 ms to 1000 ms) were labeled VSUS and TALKSPURTS longer than 10 frames (i.e. ≥ 1100 ms) were labeled NONVSUS [2, 12].

3.3. Between- and within-speaker intervals

Next, the TALKSPURTS and PAUSES of the two individual speakers were combined to identify intervals of single-speaker speech for each speaker; intervals of joint silence; and intervals of joint speech. Subsequently, between-speaker intervals and within-speaker overlaps were identified. *Between-speaker intervals* were defined as intervals of joint silence or joint speech preceded and followed by single-speaker speech by different speakers. Within-speaker overlaps were defined as joint speech preceded and followed by single-speaker speech by the same speaker.

The durations of the between-speaker intervals were calculated by subtracting the time of the offset of the preceding TALKSPURT from the onset of the nearest following TALKSPURT. This resulted in positive intervals for silences (see 1a in Figure 1) and negative intervals for overlaps (see 1b in Figure 1). For the within-speaker overlaps, we calculated the duration from the onset of joint speech to the nearest offset of single-speaker speech forward in time (see 2 in Figure 1) to obtain an interval whose duration is more readily comparable to that of between-speaker intervals.

Finally, the VSU/NONVSU distinction was used to identify four types of between-speaker intervals: VSU-VSU; VSU-NONVSU; NONVSU-VSU; and NONVSU-NONVSU.

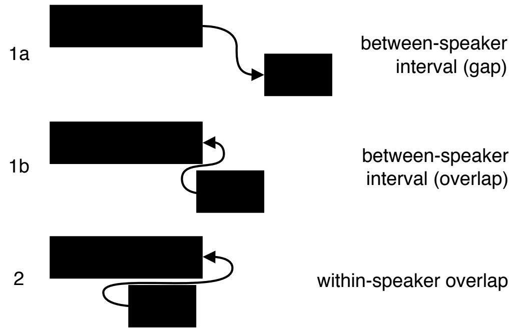

Figure 1. Illustration of duration measurements for between-speaker intervals (1a and 1b) and within-speaker overlaps (2).

4. Results

4.1. Between-speaker intervals

A total of 6506 between-speaker intervals were identified in the speech material. Table 1 presents the distribution of the four types of between-speaker intervals, as well as means, standard deviations, medians and interquartile ranges (IQR) for the durations of these. One observation from Table 1 is that the inclusion of a VSU vs. NONVSU split clearly affects the distribution. The NONVSU-NONVSU row shows lower measures of central tendency and higher variance compared to when these categories are not distinguished (cf. the Total row). The other three cells involving VSUS show higher central tendency and lower variance compared to the Total.

Table 1. Descriptive statistics for the durations (in ms) of the four types of between-speaker intervals.

	Mean	Std. dev.	Median	IQR	N
VSU-VSU	281	599	200	400	432
VSU-NONVSU	287	630	200	500	1618
NONVSU-VSU	203	480	200	300	1621
NONVSU-NONVSU	-36	832	0	700	2835
Total	125	709	100	600	6506

These between-speaker intervals were further analyzed in a two-way ANOVA with the duration of the between-speaker interval (in ms) as the dependent variable; the four types of speaker change as a fixed factor; and the 23 speaker-pairs (or dialogues in the TRAINSET) as a random factor. The ANOVA showed strong and significant effects of type of speaker change $F(3, 81) = 43.7; p < 0.001; \eta_p^2 = 0.62;$ and speaker-pair $F(22, 108) = 7.3; p < 0.001; \eta_p^2 = 0.60;$ and a weak but significant effect of the interaction between the two factors $F(66, 6414) = 1.8; p < 0.001; \eta_p^2 = 0.02.$

Thus, the speaker-pairs affected the between-speaker intervals. Furthermore, a Tukey HSD Post Hoc test on the effect of type of speaker change showed (i) that intervals delineated by two NONVSUS had significantly shorter mean durations than any of the other types; (ii) that the mean duration of the interval *before* a VSU was significantly shorter than that *following* a VSU; and (iii) that the mean duration of the interval after a VSU was not significantly affected by the type of following utterance. These findings were largely corroborated by the estimated marginal means in the interaction: (i) intervals delineated by two NONVSUS were shorter than any of the other types in 19 out of 23 speaker-pairs; (ii) NONVSU-VSU intervals were shorter than VSU-NONVSU intervals in 15 out of 23 speaker-pairs; and (iii) VSU-VSU intervals were more similar to VSU-NONVSU intervals than to any other types in 12 out of 23 speaker pairs.

As all four distributions deviate from a normal distribution (all are slightly positively skewed and leptokurtic), mean comparisons do not present the whole picture – the distribution is another relevant aspect of the timing. Figure 2 shows histograms for the four types of between-speaker intervals (bin size 200 ms), as well as the intervals between the onset of within-speaker overlap and the offset of the nearest single-speaker speech forward in time.

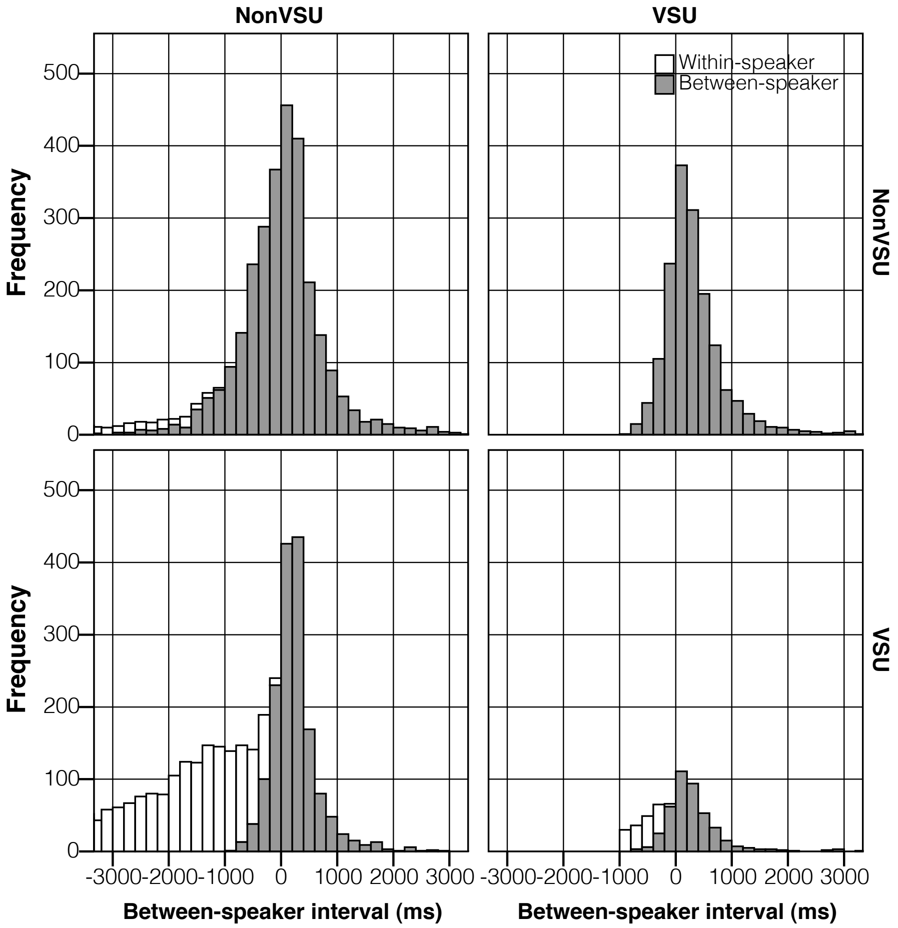

Figure 2. Histograms of between-speaker intervals in ms (grey bars) in the four types of speaker change: NONVSU -NONVSU (top left panel); VSU-NONVSU (top right panel); NONVSU-VSU (bottom left panel); and VSU-VSU (bottom right panel). The bin size is 200 ms. The white bars show the intervals from the onset of within-speaker overlap to the offset of the nearest single-speaker speech forward in time.

These histograms reveal that the between-speaker intervals in the different types of speaker changes were strikingly similar in some respects whilst differing substantially in others. For example, the most frequent between-speaker interval in all four panels are those that range between 0 and 400 ms of silence. A salient difference between the panels is that the NONVSU-NONVSUs have relatively more variance, as well as more and longer overlaps than any of the others. Conversely, between-speaker intervals delineated by VSUs on either or both sides have fewer and shorter overlaps, a finding compatible with [10].

Moreover, the observation of shorter between-speaker intervals (and smaller variance) before VSUs than after them seems warranted by the histograms (compare the NONVSU-VSU and VSU-NONVSU histograms in Figure 2). Thus, VSUs are not centered on the between-speaker interval, but rather somewhat aligned to the left.

4.2. Perceived speaker change categories

The durations of the between-speaker intervals were also related to the perceptual detection thresholds for gaps and overlaps established in [14] such that intervals of joint silence ≥ 2 frames in speaker changes were labeled GAPS; intervals of joint speech ≥ 2 frames OVERLAPS; and intervals of joint silence or speech ≤ 1 frame NO-GAP-NO-OVERLAP.

The distribution of perceived speaker change categories across the four types of between-speaker intervals (shown in Table 2) shows a similar pattern to that given by the between-speaker interval distributions. The NONVSU-NONVSUs have relatively more OVERLAPS and fewer GAPS and NO-GAP-NO-OVERLAPS than the other categories. Conversely, the types of speaker changes involving VSUs all have a larger proportion of NO-GAP-NO-OVERLAPS than the NONVSU-NON-VSUs. Furthermore, the VSU-VSUs, VSU-NONVSUs and NONVSU-VSUs are remarkably similar in the relative frequencies of the perceived speaker change categories.

Table 2. Distribution of perceived speaker change categories for the four types of speaker changes.

	Perceived as	Frequency	Percent
VSU-VSU	OVERLAP	51	11.8
	NO-GAP-NO-OVERLAP	156	36.1
	GAP	225	52.1
VSU-NONVSU	OVERLAP	224	13.8
	NO-GAP-NO-OVERLAP	551	34.1
	GAP	843	52.1
NONVSU-VSU	OVERLAP	200	12.3
	NO-GAP-NO-OVERLAP	608	37.5
	GAP	813	50.2
NONVSU-NONVSU	OVERLAP	1063	37.5
	NO-GAP-NO-OVERLAP	729	25.7
	GAP	1043	36.8

4.3. Within-speaker overlaps

A VSU that is completely overlapped by another talkspurt is not considered a between-speaker interval in the current framework, or in any existing theories of turn-taking for that matter. It is possible, however, that some proportion of the within-speaker overlapped VSUs were indeed intended as speaker changes. For a short talkspurt, the difference between being a speaker change in overlap and being a within-speaker overlap is very small: cf. panels 1b and 2 of Figure 1. The white bars in Figure 2 indicate what the distributions look like when the intervals from onset of within-speaker overlap to offset of the nearest single-speaker speech are included. Clearly, if some proportion of the short within-speaker overlaps had indeed been intended as speaker changes, and had happened to terminate just a little later, the differences between the NONVSU-NONVSU and NONVSU-VSU histograms in the bottom left and top right panels of Figure 2, respectively, would be smaller.

5. Discussion and conclusions

This study has shown that very short utterances are more precisely timed to the preceding utterance than longer utterances in terms of a smaller variance and a larger proportion of perceived no-gap-no-overlaps. Furthermore, that excluding intervals adjacent to very short utterances from distributions of between-speaker interval durations, as is sometimes done, results in measures of central tendency closer to zero as well as larger variance. These effects, however, are mainly due to a larger proportion of longer overlaps.

The observed differences between NONVSU-NONVSU and the other types of between-speaker intervals involving VSUs can probably to some extent be an artifact of the definition of VSUs in terms of duration, however. First, the maximum between-speaker overlap involving VSUs is 900 ms by definition, and although empirically overlap durations are limited by talkspurt length, there is no such theoretical upper limit on the duration of the overlaps involving two NONVSUS. This makes the variance greater in the NONVSU-NONVSUS. Second, as VSUs are short, the overlaps involving VSUs are by definition at least as short, and in practice even shorter: in fact, 95% of all overlaps involving VSUs are shorter than 500 ms. This truncates the left tail of the VSU-delineated intervals. Third, a higher number of short utterances likely results in a higher number of within-speaker overlaps in the data which do not entail speaker changes in any existing theories of turn-taking. Were they to do so, they would significantly shift the central tendency towards more negative values.

Taken together, these findings show that it is important to keep track of whether backchannels are included or not when interpreting and comparing between-speaker intervals, and that speaker changes, as traditionally defined, becomes problematic as the proportion of short utterances increase. Excluding VSU-delineated between-speaker intervals from distributions of between-speaker interval durations lowers measures of central tendency and raises variance, which in turn makes it easier to discard theories of gap and overlap minimization. Retaining VSU-delineated between-speaker intervals has the opposite effect, offering support for theories that assume that turn-taking is reactive. In this way, seemingly similar methodologies can have the perverse effect of favoring divergent interaction theories, and render it impossible to compare different studies treating different languages and/or domains.

6. Acknowledgements

We thank the three anonymous reviewers for helpful suggestions. This research was funded by the Riksbankens Jubileumsfond (RJ) project P09-0064:1-E *Prosody in conversation* and the Swedish Research Council (VR) project 2009-1766 *The rhythm of conversation*.

7. References

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

[2] Edlund, J., Heldner, M., and Pelcé, A., "Prosodic features of very short utterances in dialogue", In M. Vainio*, et al.* [Eds.], *Nordic Prosody: Proceedings of the Xth Conference, Helsinki 2008*, 57-68, Peter Lang, 2009.

[3] Heldner, M. and Edlund, J., "Pauses, gaps and overlaps in conversations", *Journal of Phonetics,* 38:555-568, 2010.

[4] ten Bosch, L., Oostdijk, N., and Boves, L., "On temporal aspects of turn taking in conversational dialogues", *Speech Communication,* 47:80-86, 2005.

[5] Weilhammer, K. and Rabold, S., "Durational aspects in turn taking", In *Proceedings of the 15th international congress of phonetic sciences (ICPhS 2003)*, 2145-2148, 2003.

[6] Yngve, V. H., "On getting a word in edgewise", In *Papers from the Sixth Regional Meeting Chicago Linguistic Society*, 567-578, Chicago Linguistic Society, 1970.

[7] Allwood, J., Nivre, J., and Ahlsén, E., "On the semantics and pragmatics of linguistic feedback", *Journal of Semantics,* 9:1-26, 1992.

[8] Clark, H. H., *Using language*, Cambridge University Press, 1996.

[9] Schegloff, E., "Discourse as an interactional achievement: Some uses of 'uh huh' and other things that come between sentences", In D. Tannen [Ed.], *Analyzing Discourse: Text and Talk*, 71-93, Georgetown University Press, 1982.

[10] Neiberg, D. and Truong, K. P., "Online detection of vocal listener responses with maximum latency constraints", In *Proceedings ICASSP 2011*, 5836-5839, 2011.

[11] Carletta, J*.*, *et al.*, "The reliability of a dialogue structure coding scheme", *Computational Linguistics,* 23:13-31, 1997.

[12] Edlund, J., Heldner, M., Al Moubayed, S., Gravano, A., and Hirschberg, J., "Very short utterances in conversation", In *Proceedings from Fonetik 2010*, 11-16, 2010.

[13] Laskowski, K., "Predicting, detecting and explaining the occurrence of vocal activity in multi-party conversation (Doctoral dissertation)," Language Technologies, Institute School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA, 2011.

[14] Heldner, M., "Detection thresholds for gaps, overlaps and no-gap-no-overlaps", *Journal of the Acoustical Society of America,* 130:2011.

[15] Edlund, J*.*, *et al.*, "Spontal: a Swedish spontaneous dialogue corpus of audio, video and motion capture", In *Proceedings of LREC 2010*, 2992-2995, 2010.