INTERSPEECH 2011 FLORENCE, ITALY

Very short utterances and timing in turn-taking

Mattias Heldner[1,2], Jens Edlund[1], Anna Hjalmarsson[1], Kornel Laskowski[1]
[1]Speech, Music and Hearing, KTH, Stockholm, Sweden
[2]New job! Department of Linguistics, Stockholm University, Stockholm, Sweden

Is there a difference in the timing of very short utterances, such as backchannels, compared to longer utterances? (i.e. timing relative to neighboring utterances)

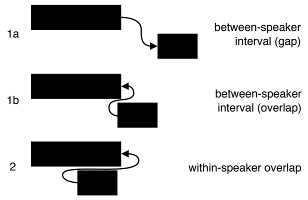
How does the inclusion/exclusion of between-speaker intervals (i.e. gaps, overlaps, no-gap-no-overlaps) adjacent to very short utterances affect the distribution of such intervals?

Speech material

A subset of the Spontal corpus, about 8 hours of spontaneous two-party face-to-face conversations in Swedish recorded with close talk microphones.

Procedure

1. Speech activity detection (100 ms frame step, 200 ms frame size) produced a segmentation into Talkspurts ≥ 200 ms; and Pauses ≥ 200 ms for each speaker.
2. The Talkspurts were subdivided into very short utterances (VSUs) and their complement (NonVSUs) based on their duration: VSUs = Talkspurts ≥ 200 ms and ≤ 1000 ms; NonVSUs = Talkspurts ≥ 1100 ms.
3. The Talkspurts and Pauses of the individual speakers were combined to identify intervals of Single-speaker speech for each speaker, Joint silence and Joint speech.
4. Sequences of such intervals were used to identify Between-speaker intervals = intervals of Joint silence or Joint speech preceded and followed by Single-speaker speech from different speakers; and Within-speaker overlaps = Joint speech preceded and followed by Single-speaker speech from the same speaker.
5. Durations of Between-speaker intervals and Within-speaker overlaps were calculated.


1a
between-speaker interval (gap)
1b
between-speaker interval (overlap)
2
within-speaker overlap

6. The VSU/NonVSU distinction was used to identify four types of between-speaker intervals: (i) VSU-VSU; (ii) VSU-NonVSU; (iii) NonVSU-VSU; and (iv) NonVSU-NonVSU.
7. Durations of Between-speaker intervals were related to detection thresholds for gaps and overlaps (Heldner, 2011): Gap = Joint silence ≥ 200 ms; Overlap = Joint speech ≥ 200 ms; No-gap-no-overlap = Joint speech or Joint silence ≤ 100 ms.

Timing differences

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

	Mean	SD	N
VSU-VSU	281	599	432
VSU-NonVSU	287	630	1618
NonVSU-VSU	203	480	1621
NonVSU-NonVSU	−36	821	2835
Total	125	709	6506

Table 2. Distribution of perceived speaker change categories for the four types of between-speaker intervals

	Perceived as	%
VSU-VSU	Overlap	11.8
	No-gap-no-overlap	36.1
	Gap	52.1
VSU-NonVSU	Overlap	13.8
	No-gap-no-overlap	34.1
	Gap	52.1
NonVSU-VSU	Overlap	12.3
	No-gap-no-overlap	37.5
	Gap	50.2
NonVSU-NonVSU	Overlap	37.5
	No-gap-no-overlap	25.7
	Gap	36.8

- Short utterances, such as backchannels, are timed differently relative to neighboring utterances than longer utterances:
- Between-speaker intervals delineated by VSUs on either or both sides have relatively fewer and shorter Overlaps and more No-gap-no-overlaps than NonVSU-NonVSUs
- Conversely, NonVSU-NonVSUs have relatively more and longer Overlaps; and fewer Gaps and No-gap-no-overlaps than the other types of between-speaker intervals

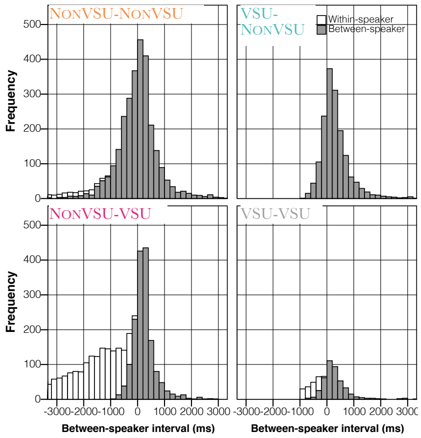


Figure 1. Histograms of between-speaker intervals (grey bars) in the four types of between-speaker intervals. The bin size is 200 ms. The white bars show the interval from the onset of Within-speaker overlap to the offset of the nearest single-speaker speech forward in time.

Distributions

- While these timing differences may to some extent be an artifact of the definition of VSUs in terms of duration, the choice to include or exclude such short utterances from distributions of between-speaker interval durations affects those distributions considerably:
- Excluding short utterances results in lower central tendency, which has been interpreted as more precise timing, but also in higher variance and relatively fewer No-gap-no-overlaps, which may well be interpreted as less precise timing
- If some proportion of the Within-speaker overlaps had indeed been intended as Overlaps,

Conclusions

It is important to keep track of whether short utterances, such as backchannels, are included or not when interpreting and comparing between-speaker intervals.

Acknowledgements

This research was funded by the Riksbankens Jubileumsfond (RJ) project *Prosody in conversation* and the Swedish Research Council (VR) project *The rhythm of conversation*.