Effects of talk-spurt silence boundary thresholds on distribution of gaps and overlaps

Marcin Włodarczak, Petra Wagner

Faculty of Linguistics and Literary Studies
Bielefeld University, Germany
{mwlodarczak,petra.wagner}@uni-bielefeld.de

Abstract

Faced with lack of objective and easily applicable criteria for segmentation of speech into dialogue turns, many authors resort instead to units defined in terms of stretches of speech minimally bounded by silence of some predefined duration. There is, however, no consensus concerning silence thresholds employed. While such thresholds can be established on perceptual grounds, in practice a wide range of values is used. As this has a direct impact on the reported frequencies of silences and overlaps, the discrepancies make comparisons of results across different studies difficult. In an attempt to overcome these problems in the present paper we use the Switchboard corpus to evaluate the expected variability in distributions of inter- and intra-speaker intervals when silence boundary thresholds of inter-pausal units are manipulated.

Index Terms: dialogue segmentation, inter-pausal units, gaps and overlaps

1. Introduction

Linguists are notorious for disagreeing about definitions of the most basic concepts of their discipline. Controversies surrounding terms such as *sentence*, *syllable* or *word* provide ample examples of differences of opinion between representatives of different theoretical schools and methodologies. Similar difficulties pertain to the definition of *dialogue turn*. The term originates in the field of Conversation Analysis (CA), which, as put by one of its proponents, "is methodologically 'impure' but [...] works" [1]. Nonetheless, owing to CA's unique methodological stance, going back to its roots in phenomenological sociology, its terms are difficult to apply in an objective and mechanistic fashion, which is a prerequisite in large-scale corpus studies.

Tellingly, the notion of turn is not formally defined in the classic formulation of a conversational turn-taking system by Sacks et al. [2]. Turns are merely stated to consist of linguistic structures which "allow a projection of the unit-type under way, and what, roughly, it will take for an instance of that unit-type to be completed." The projection in question is primarily syntactic but other factors, such as prosody, might also contribute to projection of unit endings. Points of projected completion constitute transition relevance places (TRPs), at which speaker change might occur. However, since Sacks et al.'s system mandates that a TRP might be followed by more talk from the same speaker, turn can effectively comprise many basic constituents. Neither can pausing be used to reliably delimit turn boundaries as Sacks et al. differentiate between intra-turn *pause* and inter-turn *gaps*, which, when longer than some unspecified duration, become *lapses*. What is more, the distinction between gaps and pauses is retrospectively negotiable: "if a developing silence occurs at a transition-place, and is thus a (potential) gap, it may be ended by talk of the same party who was talking before it; so the 'gap' is transformed into a 'pause' (being now intra-turn)." At the same time, as CA transcripts make clear, not all instances of silence bounded by talk from the same speaker are guaranteed to constitute intra-turn gaps as sequential and / or pragmatic criteria might warrant a segmentation into separate turns. Lastly, the concept of turn is further complicated by cases of overlapping speech, in which an attribution of turn holder is uncertain at best. Indeed, given these practical and methodological difficulties, it has been suggested that turn should be conceived of as an essentially prescriptive term [3].

Alternatively, in the tradition of Duncan and Fiske [4], turns might be defined in terms of (both verbal and non-verbal) turn-releasing and turn-taking *signals* exchanged by dialogue partners. However, as long as no clear and exhaustive specification of such signals has been arrived at, they do not provide a workable solution to dialogue segmentation.

Faced with similar difficulties a corpus linguist is likely to switch instead to a segmentation based on the purely mechanical criterion of *speaker change*. This methodological tradition has its roots in the technique of *interaction chronography*, first employed by Norwine and Murphy [5]. Interactional chronography does away with the concept of turn as its basic unit and replaces it with the notion of *talkspurt*, i.e. "speech by one party, including his pauses, which is preceded and followed, with or without intervening pauses, by speech from the other party perceptible to the one producing the talkspurt" [5, p. 282]. This terminology was further refined by Jaffe and Feldstein [6] to include unilateral *vocalisations*, *speaker switches*, *pauses* (silences bounded by vocalisation of the same speaker), *switching pauses* (silences bounded by vocalisations of different speakers) and *simultaneous speech*. This classification allowed them to model temporal patterns of speaking in dialogue stochastically as a first-order Markov model defined in terms of four *dyadic states*: unilateral vocalisation by each of the speaker, simultaneous vocalisations, and simultaneous silence.

Importantly, while the states of the model are defined fully deterministically, vocalisations and silences themselves were inferred (automatically) from signal intensity measured at intervals of predefined duration, set between 100 and 300 ms [7]. Although these thresholds are claimed to reflect perceptual constraints on silence and speech detection, the exact values used appear to have been arrived at in a fairly ad-hoc manner[1].

Since Jaffe and Feldstein's study was published, many au-

[1] In Jaffe and Feldstein's setup the thresholds were set to correspond "to the natural common sense perception of sound burst and pause in speech" [6, p. 18].

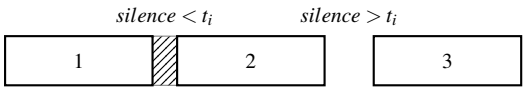

Figure 1: Construction of talkspurts given a minimum silence threshold t_i. Vocalisations 1 and 2 will be merged since silence duration between them does not exceed t_i. Vocalisations 2 and 3, which are separated by silence longer than t_i, will be left intact.

thors (including the present ones) have used talkspurts (or interpausal units) as the basic unit of dialogue segmentation, e.g. [8, 9, 10, 11, 12]. All these studies face the same problem of establishing minimal silence durations for determining vocalisation boundaries. Perceptual thresholds on perception of silences and overlaps in dialogue were investigated recently by Heldner [13] and were found to correspond to about 120 ms. In practice, however, there is no consensus concerning the choice of threshold value. While most studies set it within the range of 100-200 ms, thresholds as short as 50 ms [10] and as long as 500 ms [9] have been used. As these discrepancies are likely to result in different distributions of overlap and silence, results of studies using different silence thresholds are not directly comparable.

In addition, as demonstrated by Campione and Véronis [14], using thresholds on silence durations (both low and high) can lead to entirely wrong conclusions. When overlaps are added to the equation, the situation becomes even more complex because of interactions between the two categories. For example, vocalisations whose onsets coincide with shorter silences in interlocutor's speech will be categorised as overlapping or non-overlapping depending on the threshold value.

To at least partly overcome these problems and estimate the expected variability in distribution of silences and overlaps, we calculated mean durations and frequencies of these categories in the Switchboard corpus while manipulating the silence threshold.

2. Method

Automatically labelled phone boundaries in 642 Switchboard dialogues distributed with the NXT-format Switchboard Corpus [15] were used as basic vocalisation units. These units were then merged across a range of thresholds of minimum silence. Specifically, consecutive units were joint when silence between them was shorter than the current threshold value. This procedure is presented in Figure 1.

Threshold values ranged from 50 to 500 ms in 2.5 ms increments. For each threshold value frequencies and mean durations of the following categories were calculated:

Within-speaker silence (WSS): stretch of silence bounded by vocalisations of the same speaker.

Between-speaker silence (BSS): stretch of silence bounded by vocalisations of different speakers.

Within-speaker overlap (WSO): stretch of speech bounded by vocalisations of the same speaker.

Between-speaker overlap (BSO): stretch of speech bounded by vocalisations of different speakers.

Solo vocalisation (SOLO): unilateral vocalisation of one speaker bounded by vocalisations of the other speaker and / or silence longer than the specified duration.

These categories are portrayed in Figure 2. In addition, we identified talkspurts (SPURT) in the spirit of Jaffe and Feldstein's *possession of the floor* as the interval between the first unilateral sound of one speaker and the first unilateral sound of the other speaker.

3. Results

Mean durations and frequencies of the six interval types are plotted in Figure 3 as a function of threshold value. The curves follow the expected pattern: as increasingly long silences are bridged, mean durations of all intervals increase and their counts decrease.

The only exception are WSOs, whose count initially falls slightly, followed by a sharp rise for thresholds greater than about 150 ms. The increase might be somewhat surprising at first but is in fact a likely effect of an interaction with the BSS category: SOLO vocalisation onsets which, with smaller threshold values, coincide with short silences in interlocutor's speech, effectively classified as BSSs, are subsequently re-categorised as WSOs. Indeed, a linear regression model with BSS as an independent variable explains 84% of variance of WSO (as measured by R^2). A similar interdependency between WSOs and BSOs, most likely related to overlaps resulting from the previous speaker resuming speech after a brief pause, yields an R^2 of 0.81. This pattern, presented schematically in Figure 4, might be fairly common, and correspond for instance to feedback signals slotted into short silences in interlocutor's speech. Importantly, the increase in WSO counts occurs only for thresholds longer than roughly 150 ms, which corresponds very closely to the perceptual threshold on silence detection found by Heldner [13]. This suggests that these responses might be properly cued by silences in dialogue partner's speech, and, therefore, categorising them as WSOs is likely to be inappropriate. In other words, eliminating silences longer than 150 ms leads to overestimation of WSO counts at the expense of perceptually (and possibly functionally) salient BSSs. What is more, the original overlapper / overlappee roles are swapped as a result. Incidentally, the fact that this effect does not seem to occur for lower threshold values could indicate that, in line with earlier findings, BSSs shorter than 150 ms (i.e. responses which *are not* cued by silence) are relative rare.

Mean durations and frequencies of BSOs, BSSs and SPURTs follow essentially identical trajectories ($R^2 > 0.99$), reflecting the fact that the first two categories constitute boundary conditions of SPURT segmentation. The same is true of the WSS and SOLO categories ($R^2 > 0.99$): unilateral vocalisations are mainly extended by bridging over within-speaker silences.

While the *direction* of change in most of the categories is not particularly surprising, more relevant is its *magnitude*. To asses the latter we calculated raw and percentage change within the threshold range for each of the categories. The results are tabulated in Table 1. Both measures follow the same pattern: WSOs and BSSs exhibit the least variability (12-16% for duration, corresponding to 40-80 ms, and 17-19% for frequency). BSOs are

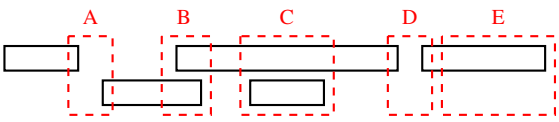

Figure 2: Categories of inter- and intra-speaker intervals used: between-speaker silence (A), between-speaker overlap (B), within-speaker overlap (C), within-speaker silence (D) and solo vocalisation (E). The top and bottom stripes represent individual speakers' vocalisations.

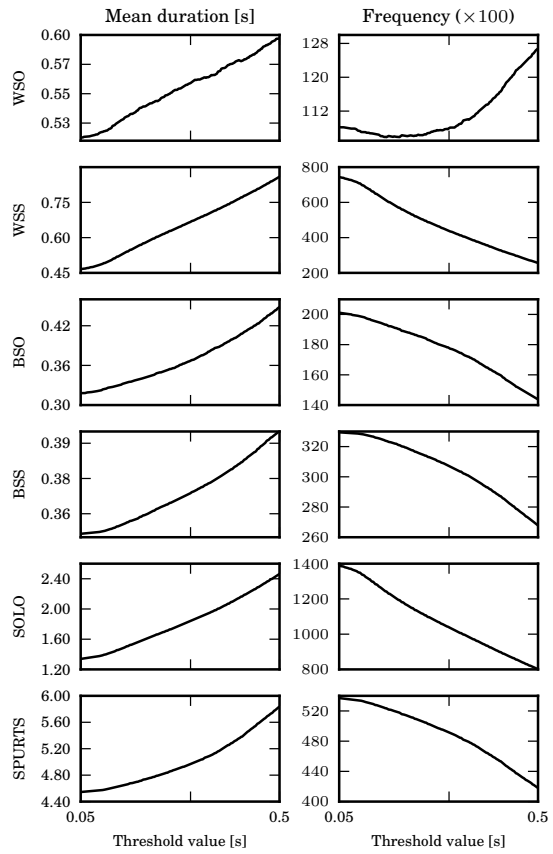

Figure 3: Mean durations (left) and frequencies (right) of within-speaker overlaps (WSO), within-speaker silences (WSS), between-speaker overlaps (BSO), between-speaker silences (BSS), solo vocalisation (SOLO), and talkspurts (SPURT) as a function of silence threshold.

much more sensitive to threshold manipulation, resulting in a 41% decrease in mean duration (130 ms) and 65% decrease in frequency. WSSs show most change of all interval types with 84% increase in duration, corresponding to 390 ms, and 65% decrease in frequency. Finally, SPURTs are more stable than SOLOs both on mean duration (28% vs. 83%) and frequency (22% vs. 42%), indicating that units defined in terms of speaker-change rather than in terms of pausing only offer a more robust segmentation baseline.

Lastly, since distributions of silences and overlaps are often characterised by ratios of frequencies in different categories (cf. [12, 9]) in Figure 5 we plot (log-transformed) ratios between

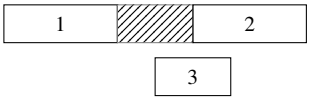

Figure 4: As vocalisations 1 and 3 are merged (represented by the shaded area) the BSO between vocalisations 1 and 2, and the BSO between vocalisations 2 and 3 are re-categorised as a single WSO, in effect obscuring the original overlappee / overlapper roles. An overlap between vocalisations 1 and 3 would have a similar effect, however without the role swapping.

Table 1: Absolute and percentage change in mean duration and frequency of the investigated categories for thresholds in the range of 0.05-0.5 s.

	Mean duration		Frequency	
	Absolute	%	Absolute	%
WSO	0.08	16	1825	17
BSO	0.13	41	−5687	-28
WSS	0.39	84	−48417	-65
BSS	0.04	12	−6097	-19
SOLO	1.12	83	−58371	-42
SPURTS	1.28	28	−11786	-22

frequencies of WSO and BSO; WSS and BSS; and between the summed frequencies of WSO and BSO, and the summed frequencies of WSS and BSS. The latter measure corresponds thus to the ratio between frequencies of all overlaps and all silences. Given the interdependencies between the categories alluded to above, these ratios are likely to exhibit significant fluctuation. This is in fact the case, especially for the WSS to BSS and the overlap to silence ratios which change more than twofold. The ratio of BSO to WSO appears to be the less affected by theshold value manipulation but still increases by more than a half (63%). Curiously, all three measures converge towards 1 (0 on log scale) around the threshold value of 500 ms.

4. Discussion and conclusions

The aim of this paper is to serve as a cautionary note about the perils of arbitrariness in selecting silence thresholds for talkspurt segmentation. The analysis outlined in the previous section leads to three basic conclusions.

First, as expected, changing the threshold value has a considerable effect on the reported durations and frequencies of between- and within-speaker categories, as well as vocalisations. Moreover, sensitivity to threshold manipulation varied across the categories, and was predictably the highest for WSSs and SOLOs, reflecting the fact that these categories are defined predominantly (or, as in the case of WSSs, entirely) in terms of pausing threshold. By contrast, categories also delimited by interlocutor's speech were relatively more robust to threshold modification but were nevertheless substantially influenced by it.

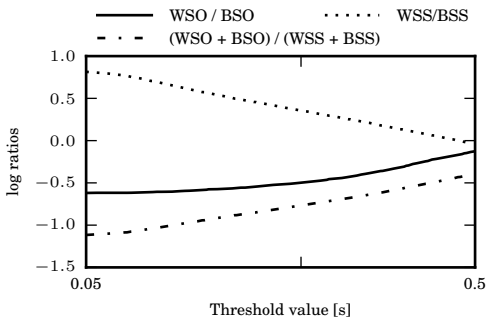

Figure 5: Log ratios of interval frequencies for thresholds in the range of 0.05-0.5 s

Second, because of the interdependencies between the categories, changing the boundary conditions leads to rellocation of instances between the categories. This is potentially dangerous and might have unexpected consequences for the obtained results. In particular, we observed that threshold values significantly exceeding the perceptual threshold on pause detection might result in overestimation of WSOs frequencies, potentially obscuring interactionally salient phenomena such as feedback-cueing pauses, and leading to reassignment of overlapper / overlappee roles.

Third, reporting frequency ratios rather than raw counts does not in itself gurarantee stability of results across threshold values. Indeed, ratios may as much as double (or halve) within the investigated threshold range.

More generally, the technique used in the present paper offers an alternative way of analysing distributions of events in dialogue, one which accounts for interdependencies and interactions between them. We plan to pursue this line of inquiry in the future.

5. Acknowledgements

This work was in part supported by German BMBF-funded "Professorinnenprogramm" FKZ 01FP09105A grant.

6. References

[1] P. ten Have, "Methodological issues in conversation analysis," *Bulletin de Méthodologie Sociologique*, vol. 27, no. 1, pp. 23–51, 1990.

[2] H. Sacks, E. A. Schegloff, and G. Jefferson, "A simplest systematics for the organization of turn-taking for conversation," *Langauge*, vol. 50, no. 4, pp. 696–735, 1974.

[3] D. C. O'Connell, S. Kowal, and E. Kaltenbacher, "Turn-taking: A critical analysis of the research tradition," *Journal of psycholinguistic research*, vol. 19, no. 6, pp. 345–373, 1990.

[4] S. Duncan and D. W. Fiske, *Face-to-face interaction: Research, methods, and theory*. Hillsdale, NJ: Erlbaum, 1977.

[5] A. C. Norwine and O. J. Murphy, "Characteristic time intervals in telephonic conversation," *Bell System Technical Journal*, vol. 17, no. 2, pp. 281–291, 1938.

[6] J. Jaffe and S. Feldstein, *Rhythms of dialogue*. New York: Academic Press, 1970.

[7] S. Feldstein and J. Welkowitz, "A chronography of conversation: In defence of an objective approach," in *Noverbal behavior and communication*, 2nd ed., A. W. Siegman and S. Feldstein, Eds. Hillsdale, NJ: Erlbaum, 1987, pp. 435–499.

[8] H. Koiso, Y. Horiuchi, S. Tutiya, A. Ichikawa, and Y. Den, "An analysis of turn-taking and backchannels based on prosodic and syntactic features in japanese map task dialogs," *Language and Speech*, vol. 41, no. 3/4, pp. 295–321, 1998.

[9] E. Shriberg, A. Stolcke, and D. Baron, "Observations on overlap: Findngs and implications for automatic processing of multi-party conversation," in *Proceedings of EUROSPEECH*, 2001, pp. 1359–1362.

[10] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech and Language*, vol. 25, no. 3, pp. 601–634, 2011.

[11] M. Heldner, J. Edlund, A. Hjalmarsson, and K. Laskowski, "Very short utterances and timing in turn-taking," in *Proceedings of Interspeech 2011*, 2011, pp. 2837–2840.

[12] M. Heldner and J. Edlund, "Pauses, gaps and overlaps in conversations," *Journal of Phonetics*, vol. 30, no. 4, pp. 555–568, 2010.

[13] M. Heldner, "Detection thresholds for gaps, overlaps, and no-gap-no-overlaps," *Journal of Acoustical Society of America*, vol. 130, no. 1, pp. 508–513, 2011.

[14] E. Campione and J. Véronis, "A large-scale multilingual study of silent pause duration," in *Proceedings of Speech Prosody*, Aix-en-Provence, France, 2002, pp. 199–202.

[15] S. Calhoun, J. Carletta, J. Brenier, N. Mayo, D. Jurafsky, M. Steedman, and D. Beaver, "The NXT-format Switchboard corpus: A rich resource for investigating the syntax, semantics, pragmatics and prosody of dialogue." *Language Resources and Evaluation*, vol. 44, no. 4, pp. 387–419, 2010.