Syllable boundary effect: temporal entrainment in overlapped speech

Marcin Włodarczak[1], *Juraj Šimko*[1,2], *Petra Wagner*[1]

[1]Faculty of Linguistics and Literary Studies, [2]CITEC
Bielefeld University, Bielefeld, Germany
{mwlodarczak,juraj.simko,petra.wagner}@uni-bielefeld.de

Abstract

In this paper we investigate one aspect of temporal entrainment in dialogue, namely how onsets of overlapped speech are timed with respect to syllable boundaries of the current speaker. Overlap initiation were found to be more frequent around syllable boundaries than at other locations within the syllable. Some evidence that the turn onset timing is also influenced by the regularity of the preceding syllables and by changes in speech tempo is also presented.

Index Terms: dialogue rhythm, temporal entrainment, overlapped speech, turn-taking

1. Introduction

Dialogue participants mutually influence various prosodic and temporal characteristics of each other's speech. It has been demonstrated that dialogue partners' utterances converge in a number of prosodic features like F0, intensity, voice quality and speech rate [1, 2, 3, 4]. This convergence has been discussed under various labels such as alignment, entrainment or mimicry.

A similar kind of entrainment has been suggested for *temporal* and *rhythmic* characteristic of speech. Several accounts of temporal dependency between subsequent utterances have been presented, mostly in the context of turn taking. These studies focus on how can an identifiable temporal landmark in a turn, e.g. a syllable boundary, be aligned – in time – with similar landmarks of the previous speech act of the dialogue partner. For example, Couper-Kuhlen's model [5], based on the perceptual speech isochrony, predicts that the first accented syllable of the next turn coincides with the extrapolated sequence of accented syllables of the previous turn. The adequacy of the model has been questioned by [6] due to lack of empirical evidence.

Similarly, Wilson and Wilson proposed a model in which the likelihood of turn initiations is controlled by an oscillatory function with the frequency of oscillations determined by speaker's syllable rate [7]. In order to minimise simultaneous starts, the listener's oscillator is counterphased to that of the speaker. The choice of syllables as the underlying unit has been suggested by Wilson and Zimmermann's finding that between-speaker intervals tend to be multiples of a fixed duration [8]. Although this duration varies from conversation to conversation, its range of 80–180 ms and average duration of 120 ms matches roughly the duration of a single syllable. Interestingly, this fixed duration also corresponds to the *theta* frequency range of endogenous oscillators in the human brain, which could provide the neural mechanism for turn-taking.

Since these models have been put forward to explain the smoothness and precision of turn-taking, they intentionally avoid the phenomenon of *overlapping speech*. Indeed, Wilson and Wilson admit that "a full picture of the timing of turn-taking would require some knowledge of the durations of overlaps" [7]. This need is particularly pressing in the light of the recent re-evaluation of the amount of overlapping speech in natural conversation. In contrast to the assumption that dialogue participants are trying to minimise gaps and overlaps between turns [9], durations of between-speaker intervals found in dialogue corpora suggest that this assumption might in fact be wrong. For example, a study based on three languages (Dutch, Swedish and Scottish English) shown that smooth speaker changes constitute only a small proportion of all speaker changes and and overlaps made up to 40% of inter-speaker intervals [10]. Other reports of the proportion of overlapping turn changes found in literature vary from about 5% [11] to over 50% [12]. There is even a report of dialogue participants spending more than half of their *speaking time* in overlap [13].

The timing of overlap initiations has been studied, e.g., by Jefferson [14]. While claiming that "there is no point in an utterance which is proof from systematically accountable (if not interactionally legitimate) overlap" she admitted that "[i]n the apparent chaos of overlapping talk one can begin to locate a series of 'fixed points' which collect and order an enormous amount of [overlapping] talk", and went on to distinguish between three overlap onset types: transitional (a side-effect of imprecise turn-taking), recognitional (occurring when the overlapper has understood the thrust of the talk in progress prior to its completion point) and progressional (triggered by disfluencies or other production problem of the original speaker).

Our goal is similar to that of Jefferson. We investigate the likelihood of overlap initiation attributable to various temporal landmarks in ongoing talk. We, however, choose a much finer time scale of a *syllable* rather than that of a turn. In this study we present and analyze an evidence that the timing of initiation of an overlapping turn is linked to the syllable boundaries in the overlapped speech act of dialogue partner. In short, our results suggest a tendency towards (in-phase) temporal alignment of syllable boundaries in the mutually overlapping utterances. They are thus broadly compatible with oscillatory accounts of turn taking discussed above.

2. Method

We used the Switchboard-1 Release 2 corpus [15]. Stretches of overlapping speech were calculated from *MS-State* word-alignments [16] concatenated into inter-pausal units (IPUs) bounded by at least 100 ms of silence. For each overlap, the first overlapped syllable of the overlappee's IPU (i.e., the syllable during which the overlap was initiated) was identified. The overlap onset was then normalized relative to the duration of this first overlapped syllable: the *syllable-normalized onset time* was calculated by dividing the duration of the interval from the onset

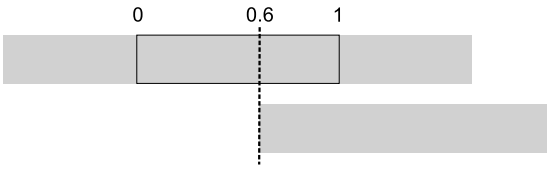

Figure 1: Overlap onset relative to the duration of the coinciding syllable in overlappee's speech. The top stripe represents overlappee's speech, 0 and 1 mark the boundaries of the overlapped syllable. The bottom stripe represents the onset of the overlapping speech.

of the overlapped syllable to the onset of the overlapping utterance by the duration of the overlapped syllable. The procedure is illustrated in Figure 1.

Overlaps coinciding with overlappee's IPU-initial and IPU-final syllables were excluded from the analysis with a view to eliminating simultaneous starts (whose timing could be expected to be random) and terminal overlaps (which are related to predicting utterance boundaries rather than syllable boundaries). Since automatically derived syllable boundaries included in the NXT Switchboard distribution [17] were used, excluding simultaneous starts from the analysis has the additional advantage of avoiding possible segmentation errors due to the speech signal starting simultaneously in both channels. Overall 10274 overlaps were analysed.

In order to verify whether the precision of temporal alignment depends on the regularity of syllable durations, for 7890 overlaps preceded by at least three syllables of overlappee's speech, normalised Pairwise Variability Index (nPVI) of those syllables and the overlapped syllable was calculated. nPVI [18] measures how much neighbouring syllables differ in duration and normalises for differences in tempo, with low values indicating high regularity of intervals:

$$nPVI = 100 \times \left[\sum_{k=1}^{m-1} \left| \frac{d_k - d_{k+1}}{(d_k + d_{k+1})/2} \right| /(m-1) \right]$$

where d_i is the duration of the i-th interval. Additionally, we calculated nPVI of those syllables ignoring the absolute value. This way we are able to capture whether the syllable durations are on average increasing (negative values) or decreasing (positive values). We refer to that metric as *directional PVI* (dPVI).

3. Results

Figure 2 shows a histogram of all syllable-normalized onset times in the corpus. One sample Kolmogorov-Smirnov test (referred to as 1-KST in the remaining of this paper) was used to verify whether these data correspond to a uniform (flat) distribution. The null hypothesis was rejected with a p-value < 0.001, i.e., the observed non-flatness of the plotted histogram is significant. The bimodality of the distribution with peaks at 0 and 1 indicates that the likelihood of an overlap initiation is higher around syllable boundaries than in the middle of a syllable. The fact that the peak at 1 is somewhat higher than that at 0 might suggest that barging in speakers "aim" at a time point just before interlocutor's syllable beginnings.

This result suggests that the turn initiation time of the barging-in speaker is influenced by the syllable boundaries of the overlappee's speech. If this is the case, it is reasonable to expect that the shape of the distribution of the syllable-normalized onset

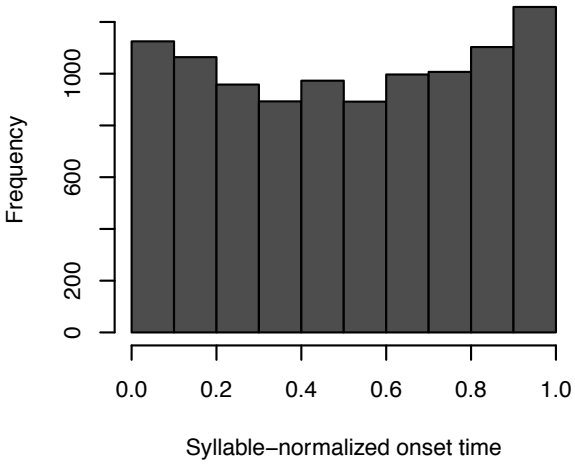

Figure 2: Distribution of syllable-normalised onset times.

times reflects the temporal structure of syllables immediately preceding the incoming partner's turn initiation. In particular, regularly spaced syllable boundaries might provide a better guidance for turn initiation and thus lead to more prominent curvature of the distribution.

In order to test this hypothesis, we used nPVI measure as an estimate of the syllabic regularity. Syllable-normalized onset times of overlaps preceded by at least three syllables of overlappee's speech were split into three equally sized classes depending on nPVI values calculated for those syllables (using 33% and 66% percentiles of the nPVI values). The resulting histograms are presented in Figure 3. Each of these distributions was significantly different from the uniform distribution (1-KST, $p < 0.001$ for all nPVI classes). Additionally, the distributions were compared pairwise by means of a two-sample Kolmogorov-Smirnov test (2-KST). Of the three possible combinations, only the classes with the highest and the lowest nPVI values differed significantly from each other ($p < 0.05$).

The high and low nPVI are classes are indeed noticeably different. The distribution of the low class is unimodal with the peak at 1, whereas the distribution of the high class is bimodal with peaks around 0 and 1. In other words, if the preceding syllables are regular, incoming speakers tend to start speaking towards syllable ends, i.e., slightly *before* the onset of the following syllable in the partner's speech. By contrast, irregularity of the temporal structure blurs this trend somehow. The less prominent peak at 1 and another peak around 0 are consistent with lower degree of precision of turn initiation towards syllable ends.

This tentative interpretation is supported by the results presented in Figure 4, where overlap onsets were split into two classes depending on the sign of dPVI, thus separating contexts in which the syllable durations are decreasing (top) from those in which they are increasing (bottom). Both of these distributions are significantly different from the uniform distribution (1-KST, $p < 0.001$). While they are not significantly different from each other (1-KST, $p = 0.45$), they are in line with the interpretation outlined above. Specifically, the peak at 0 in the top histogram (absent in the low nPVI class in Figure 3) might indeed represent overlap initiations targeted at the previous syllable which ended early. Somewhat surprisingly, however, the peak at 1 is still present there as well. Similarly, in the bottom histogram, onsets

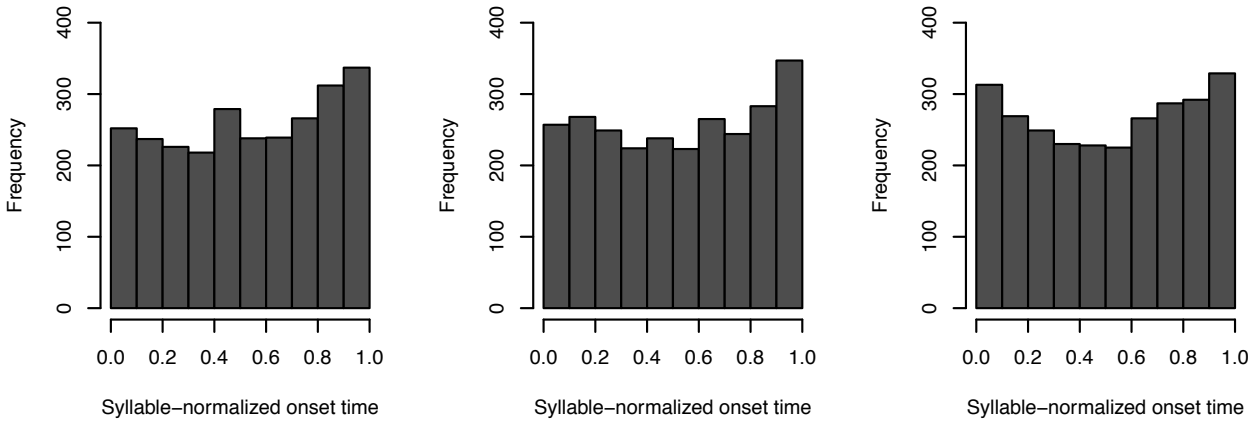

Figure 3: Distributions of overlap onsets for low (left), mid (middle) and high (right) nPVI classes of three syllables preceding an overlap.

falling early within the lengthened syllable flatten out the slope of the low nPVI class in Figure 3.

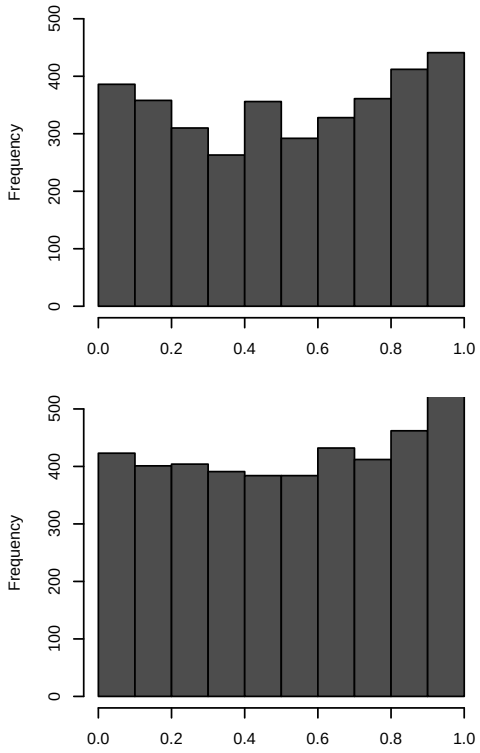

Figure 4: Distributions of syllable-normalised onset times for positive (top) and negative (bottom) dPVI classes of three syllables preceding an overlap.

The presence of measurable fine-grained temporal alignment between dialogue partners is somewhat surprising. Therefore, we paid a considerable attention to examining possible sources of the presented evidence that might be a by-product of the technical processing of the analyzed material, e.g., an impact of crosstalk between channels on automatic syllable segmentation. In order to evaluate the crosstalk effects, 1732 syllable-normalized overlap onsets were also calculated for a subset of the Switchboard dialogues marked with the lowest interchannel crosstalk values [19]. This distribution was found not to be significantly different from the overall distribution (2-KST, $p = 0.19$), nor from the distribution of relative overlap onsets calculated for Switchboard dialogues with higher crosstalk values (2-KST, $p = 0.07$). However, it was also found not significantly different from a uniform distribution (1-KST, $p = 0.48$). However, as can be seen in Figure 5, in which these overlaps are plotted against the overall distribution, overlaps from these dialogues make up only a small fraction of all overlaps. All this suggests that the absence of statistically significant curvature of the distribution might be an effect of data sparsity. To verify it 10,000 random samples of the same size (1732) were drawn from the overall distribution. As many as one in four of those samples was not significantly different from the uniform distribution.

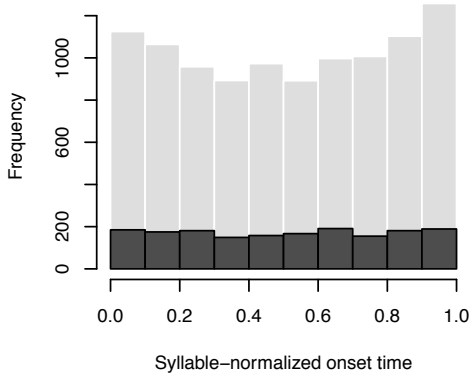

Figure 5: Distribution of syllable-normalised onset times of Switchboard dialogues with minimal amount of crosstalk (dark bars) plotted against the overall distribution (light bars).

4. Discussion

The results presented in the previous section suggest that dialogue participants indeed coordinate overlap onsets with their interlocutor's syllable boundaries. Specifically, overlappers seem to able to tune in to the current speaker's syllable rate and attempt to time the beginning of their speech with respect to a

syllable boundary in their interlocutor's speech.

The results shown in Figures 3 and 4 indicate that the coordination is at least to some extent dependent on the regularity of the preceding syllables as measured by nPVI values. However, it should be noted that nPVI is a very crude measure of regularity in speech and speakers could be expected to have mechanisms capable of capturing more complex temporal patterns than pairwise comparisons of segment durations. Therefore, clearer patterns could be expected when using more sophisticated models.

The temporal scope of temporal coordination suggested by these results might seem initially surprising. However, it has been demonstrated that in certain conditions humans are capable of an almost perfect synchronization. For example, Cummins found that speakers instructed to read a text in synchrony achieve asynchronies as small as 40-60 ms [20]. Arguably, the task of predicting a syllable boundary in spontaneous speech is a much more difficult one but it could be expected to use the same underlying cognitive mechanism.

Overall, these results are broadly compatible with rhythm-based models of turn-taking. In particular, coupled oscillator approach offers itself as a suitable starting point for modelling these phenomena. However, the shape of the overall distribution in Figure 2 and of the low nPVI class in Figure 3 indicate that listeners target syllable boundaries rather than mid points of syllables, as suggested by [7]. In terms of the oscillatory mechanism assumed in that model, speaker's and listener's oscillators seem to be more or less in phase, rather than being shifted by half a period. Although a small peak is visible both in the overall distribution and in the low nPVI class around the value of 0.5 (indicating a presence of counter-phase alignment), it could be merely an artifact of the data. At any rate, it is nowhere near as prominent as the maxima around the syllable boundaries.

The results presented here demand further attention. In order to find out what mechanisms are at play here, more extensive analysis is required in terms of speech material and rhythmic units. The results must be tested for other speech corpora, and preferably for various languages. Also, other suprasegmental units, e.g., interstress intervals, should be analyzed in this manner.

These enquiries will lead to refinements of modelling effort broadly outlined in this paper. Understanding the human ability to initiate a turn – overlapped or not – at appropriate times is not only of utmost interest theoretically, it is also one of the prerequisites for human-like mixed-initiative dialogue systems [21].

5. Acknowledgements

This research was partly supported by Alexander von Humboldt Fellowship grant to the second author.

6. References

[1] J. S. Pardo, "On phonetic convergence during conversational interaction," *Journal of the Acoustical Society of America*, vol. 119, no. 4, pp. 2382–2393, 2006.

[2] R. Levitan and J. Hirschberg, "Measuring acoustic-prosodic entrainment with respect to multiple levels and dimensions," in *Proceedings of Interspeech 2011*, Florence, Italy, August 2011, pp. 3081–3084.

[3] I. Finlayson, M. Corley, and R. Lickley, "Alignment in rate of speech: Evidence from a corpus of dialogue," in *Book of Abstracts of AMLaP 2011*, Paris, 2011, p. 67.

[4] R. L. Street, Jr., "Speech convergence and speech evalutaion in fact-finding interviews." *Human Communication Research*, vol. 11, no. 2, pp. 139–169, 1984.

[5] E. Couper-Kuhlen, *English speech rhythm: form and function in everyday verbal interactions.* Amsterdam: John Benjamins, 1993.

[6] M. Bull, "An analysis of between-speaker intervals," in *Proceedings of the Edinburgh Linguistics Conference '96*, Edinburg, 1996, pp. 18–27.

[7] M. Wilson and T. P. Wilson, "An oscillator model of the timing of turn taking," *Psychonomic Bulletin and Review*, vol. 12, no. 6, pp. 957–968, 2005.

[8] T. P. Wilson and D. H. Zimmerman, "The structure of silence between turns in two-party conversation," *Discourse Processes*, vol. 9, no. 4, pp. 375–390, 1986.

[9] 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.

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

[11] S. C. Levinson, *Pragmatics*. Cambridge: Cambridge University Press, 1983.

[12] L. ten Bosch, N. Oostdijk, and L. Boves, "On temporal aspects of turn taking in conversational dialogues," *Speech Communication*, vol. 47, no. 1–2, pp. 80–86, 2005.

[13] N. Campbell, "Approaches to convesational speech rhythm: speech activity in two-person telephone dialoges," in *Proceedings of ICPhS XVI*, Saarbrucken, 2007, pp. 343–348.

[14] G. Jefferson, "Notes on some orderlinesses of overlap onset," in *Discourse analysis and natural rhetoric*, V. D'Urso and P. Leonardi, Eds. Cleup Editore, 1984, pp. 11–38.

[15] J. J. Godfrey, E. Holliman, and J. McDaniel, "SWITCHBOARD: Telephone speech corpus for research and development," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, San Francisco, CA, 1992, pp. 517–520.

[16] N. Deshmukh, A. Ganapathiraju, A. Gleeson, J. Hamaker, and J. Picone, "Resegmentation of switchboard," in *Proceedings of ICSLP*, Sidney, Australia, 1998, pp. 1543–1546.

[17] 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.

[18] E. Low and E. Grabe, "Prosodic patterns in Singapore English," in *Proceedings of the XIIth ICPhS*, vol. 3, Stockholm, Sweden, 1995, pp. 636–639.

[19] *SWITCHBOARD: A user's manual*, http://www.ldc.upenn.edu/Catalog/readme_files/switchboard.readme.html. Accessed 22.11.11.

[20] F. Cummins, "On synchronous speech," *Acoustic Research Letters Online*, vol. 3, no. 1, pp. 7–11, 2002.

[21] J.-i. Hirasawa, M. Nakano, T. Kawabata, and K. Aikawa, "Effects of system barge-in responses on user impressions," in *Sixth European Conference on Speech Communication and Technology*, vol. 3, 1999, pp. 1391–1394.