ISCA Archive
http://www.isca-speech.org/archive

INTERSPEECH 2012
ISCA's 13th Annual Conference
Portland, OR, USA
September 9-13, 2012

# Temporal entrainment in overlapped speech: Cross-linguistic study

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

[1]Faculty of Linguistics and Literary Studies, [2]CITEC
Bielefeld University, Germany

{mwlodarczak,juraj.simko,petra.wagner}@uni-bielefeld.de

10.21437/Interspeech.2012-186

## Abstract

In a previous paper we investigated how onsets of overlapped speech in English are timed with respect to syllable boundaries of the current speaker [1]. Overlap initiations were found to be more frequent around syllable boundaries than at other locations within the syllable. In this paper we extend the previous analysis by reporting on results from two other corpora in two different languages (French and German). We found similar trends in all three datasets with an increased likelihood of an overlap initiation shortly before vowel onsets in the interlocutor's speech.

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

## 1. Introduction

When taking part in a conversation, dialogue partners influence each other's speech characteristics in many subtle ways. Inter-speaker convergence (also referred to as alignment or mimicry) has been demonstrated for prosodic features such as F0, intensity, voice quality and speaking rate [2, 3, 4]. In this paper we focus on another aspect of such mutual effect, namely, the details of temporal and rhythmic entrainment between dialogue partners.

Studies of temporal dependency between interlocutors have been conducted mainly in the context of smooth *turn taking*. Models of temporal alignment between interlocutors focus on temporal coincidence of speech landmarks–syllable or foot boundaries–in subsequent turns.

Based on the concept of perceptual isochrony, Couper-Kuhlen's model [5], for example, predicts that the first accented syllable of a turn temporally coincides with the extrapolated sequence of accented syllables of the previous turn of the dialogue partner. In a similar model proposed by Wilson and Wilson [6], the likelihood of turn initiations is controlled by an oscillatory function with frequency of oscillations determined by speaker's syllable rate. Listener's oscillator determining the onset of the turn is counter-phased to that of the speaker in order to avoid simultaneous starts. Wilson and Wilson have chosen syllables as the underlying unit based on the finding that between-speaker intervals tend to be multiples of a fixed duration similar to that of a single syllable [7].

In general, however, the approaches to date lack a solid support of empirical evidence [8]. Recently, Beňuš evaluated a number of predictions of Wilson and Wilson's model using a corpus of English task-oriented dialogues [9]. His findings offered but a weak support for the model; moreover, contrary to the model assumptions, it was pitch accents rather syllables that provided slightly better alignment between subsequent turns.

One of the possible reasons behind these gaps in supporting evidence is the explicit focus of these models on smooth turn-taking. *Overlapping speech*, a natural companion of turn-taking in spontaneous interaction, has not been sufficiently addressed in this context (neither theoretically, nor experimentally), presumably because it has long been seen as a relatively infrequent phenomenon. However, contrary to the assumption that dialogue participants are trying to minimize gaps and overlaps between turns [10], a study based on three languages (Dutch, Swedish and Scottish English) revealed that overlaps made up to 40% of inter-speaker intervals [11]. Other reports of the proportion of overlapping turn changes found in literature

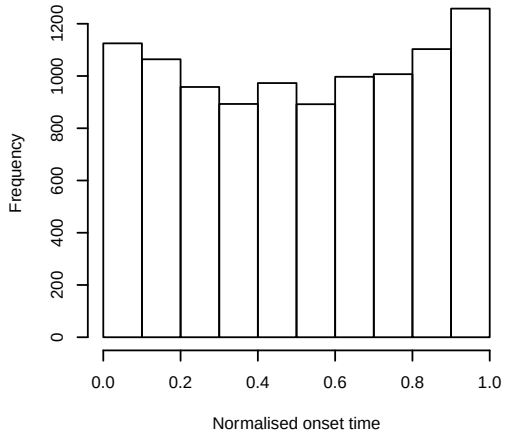


Figure 1: Distribution of onset times normalised to the duration of the first overlapped syllable. Reproduced from [1], see Section 2 for an explanation.

vary from about 5% [12] to over 50% [13]. In fact, dialogue partners may spend more than half of their *speaking time* in overlap [14].

In a previous paper [1], we investigated how onsets of overlapped speech in US English are timed with respect to syllable boundaries of the current speaker. Overlap initiations were found to be more frequent around syllable boundaries than at other locations within the syllable (see Figure 1 and Section 2 for an explanation of the method). Additionally, regularity of the preceding speech (measured with normalised Pairwise Variability Index) was found to have an effect on the timing of overlaps: an increased likelihood of an overlap before the syllable offset was observed if preceding syllables displayed a regular pattern. Changes in speaking rate were also found to play some role in coordination patterns depending on whether durations of preceding syllables were on average increasing or decreasing. Further evidence of inter-speaker entrainment was found for inter-stress intervals in the same data set [15]. A tendency was observed for overlap onsets to occur in the middle of the interval between successive stressed vowels in interlocutor's speech.

While we interpreted these findings as evidence that speakers indeed coordinate their turn onsets with the structure of their partner's speech in a manner broadly consistent with the turn-taking accounts discussed above, we were aware of several possible objections to our approach. First, as the results were based on analysis of a single speech corpus of US English, cross-language generalisations were impossible. Second, the syllables in the analysed corpus were identified using an automatic forced alignment method which might be prone to errors caused by varying degrees of cross-talk.

In this paper we thus extend the previous analysis to two other corpora in two different languages (French and German), with one corpus (German) labelled manually.

## 2. Method

The corpora used were: the Switchboard corpus [16] for English, the CID corpus [17, sldr000720] for French and the Kiel corpus of spontaneous speech (the "Lindenstrasse corpus") [18] for German. All corpora consisted of spontaneous, dyadic dialogues. For English and German stretches of overlapping speech were derived automatically from inter-pausal units (IPUs) bounded by at least 100 ms of silence. For French the IPUs distributed with the corpus (excluding units consisting solely of non-verbal phenomena, such as laughter) were used for calculating overlaps. Since no syllable segmentations were available for the Kiel corpus, we could not replicate the syllable-based findings of [1] for German. Therefore, intervals between consecutive vowel onsets (henceforth vowel-to-vowel intervals, VTV) were calculated for all three corpora, and these were used for evaluation of the degree of temporal coordination between dialogue partners.

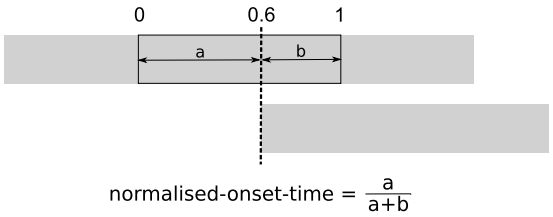


Figure 2: Overlap onset relative to the duration of the coinciding vowel-to-vowel interval in overlappee's speech. The top stripe represents overlappee's speech, 0 and 1 mark the boundaries of the overlapped vowel-to-vowel interval. The bottom stripe represents the onset of the overlapping speech.

For each overlap, the first overlapped VTV of the overlappee's IPU (i.e., the VTV during which the overlap was initiated) was identified. The overlap onset was then normalised relative to the duration of this first interval: the *normalised onset time* was calculated by dividing the duration of the interval from the onset of the overlapped VTV to the onset of the overlapping utterance by the duration of the overlapped VTV. The procedure is illustrated in Figure 2. (Results shown in Figure 1 were obtained by an analogous method using syllables instead of VTV intervals.)

Overlaps coinciding with overlappee's IPU-initial and IPU-final VTVs 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). Overall, 7697 overlaps were analysed for English, 2362 for French and 260 for German.

## 3. Results

Histograms of normalised onset times for the three corpora are plotted in Figure 3. One-sample Kolmogorov-Smirnov test was used to verify whether these data correspond to a uniform (flat) distribution. The null hypothesis stating that each of the samples was drawn from a uniform distribution was rejected at $p < 0.001$ for English and French and at $p < 0.05$ for German. The significant result is particularly notable for the Kiel corpus given the small data set and the manual phone segmentation, which alleviates the concerns about the result being an artefact introduced by forced alignment discussed in [1].

While the pattern is somewhat less easily discernible in the CID corpus, in all three corpora the occurrence of an overlap initiation rises while moving away from the vowel onset and reaches a maximum at around 80% of the overlapped VTV interval (that is shortly before the next vowel onset) followed by a small fall. Although the distributions' shapes are visually different in some details (e.g., the slow and spread-out rise in the CID corpus versus

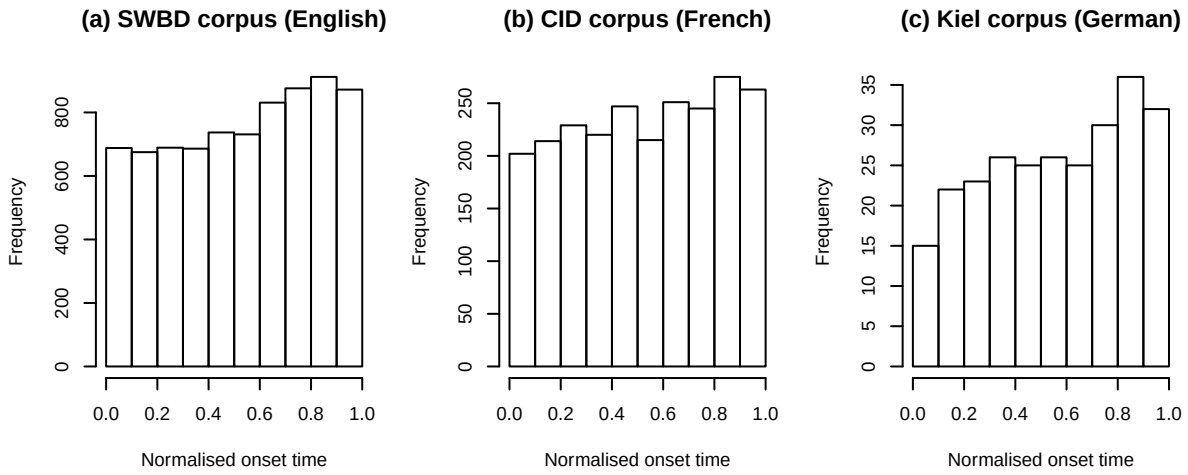


Figure 3: Distributions of normalised onset times for (a) the Switchboard corpus, (b) the CID corpus, (c) the Kiel Corpus of Spontaneous Speech.

the late and much steeper rise in Switchboard, and two rises at the beginning and end separated a by a plateau in the Kiel corpus), the differences are not statistically significant (two-sample Kolmogorov-Smirnov test, $p > 0.05$ for all pairwise comparisons).

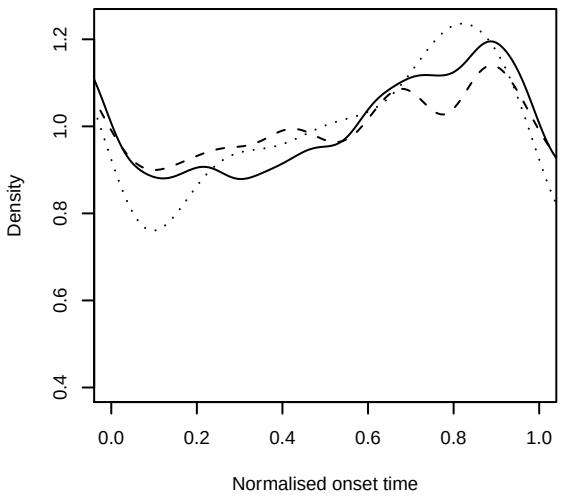


Figure 4: Density functions of normalised onset times for the three corpora. The Switchboard corpus: solid line, the CID corpus: dashed line, the Kiel corpus: dotted line.

To allow for a better comparison of the corpora, density functions of all three distributions were calculated. To ensure that density values do not drop around the edges each distribution was joined with its copies on the right and on the left before calculating the density functions. This is desirable because of the cyclical nature of the phenomenon in question. The resulting functions, plotted in Figure 4, further emphasise the similarities between the normalised onset time distributions with a rising trend towards the following vowel boundary.

Although we could not use syllable boundaries on the German material, onset times normalised with respect to syllable boundaries were calculated for the French CID corpus using a procedure similar to that in [1] (results for the US English Switchboard corpus are shown in Figure 1). The normalised onset time histogram computed using the automatic syllable segmentations distributed with the corpus was not significantly different from a flat distribution (one-sample Kolmogorov-Smirnov test, $p > 0.05$).

## 4. Discussion

Our findings present a solid evidence for temporal entrainment between dialogue partners. They corroborate the previously reported results for English: the change of the reference frame–from syllables to VTV intervals–did not obliterate the evidence for tendency of dialogue partners to start their turn at a particular juncture (phase) within syllable-sized units. As shown in Figure 1 and the leftmost panel of Figure 3, the shift of temporal reference in time to the right (syllable onsets precede the onset of syllable nuclei) resulted in the expected shift of the (in principle circular) distribution of normalised onset times to the right.

Interestingly, for the French data the evidence for interspeaker coordination is found on the basis of vocalic onsets, but *not* for syllable onsets. If the reported tendencies are indeed explicable in terms of rhythmic entrainment to the conversational partner's speech, this result could perhaps be accounted for by the perceptually (rather then

phonologically) dominant speech events referred to as perceptual centres. The perceptual centre–or p-centre–is postulated to be the instant of the *perceived* syllable onset and often corresponds to the vocalic onset [19]. The p-centre has also been found to be the temporal anchor of listeners' when asked to tap along to speech [20] or when asked to synchronise speech and auditory pulses [21]. Our results can thus be interpreted as providing additional evidence for the detected tendencies being the product of a temporal entrainment process guided by interlocutors' perception of rhythmic characteristics of each other's speech. In other words, the reported findings are compatible with the previously observed tendency of speakers to temporally entrain to each other's syllable onsets, here estimated using the more perceptually salient p-centres (vowel onsets).

The density estimates of normalised onset times shown in Figure 4 are essentially identical for all three languages, all peaking around 80% of the normalised VTV interval. This suggests that the postulated inter-speaker entrainment is not language dependent, but is grounded in a more universal properties of the speech perception–production apparatus. Moreover, the results alleviate our concerns about the identified tendency being a mere by-product of a possible interference of cross-talk with the automatic syllable labelling procedure used for the Switchboard and CID corpora. In fact, the shape of the estimated density function is the most pronounced–at least visually–for the German Kiel corpus for which phones were labelled manually. This is even more surprising given the relatively small size of the analysed German material compared to the other two languages.

Although the evidence presented here was identified using overlapped speech, we believe that the results can inform research on turn-taking and inter-speaker coordination in general. Specifically, our findings might be used to partly fill in the missing empirical evidence much needed for refining the models of temporal alignment between interlocutors.

## 5. Acknowledgements

This work was in part supported by German BMBF-funded "Professorinnenprogramm" FKZ 01FP09105A grant to the first author and by the von Humboldt Fellowship grant to the second author.

## 6. References

[1] M. Włodarczak, J. Šimko, and P. Wagner, "Syllable boundary effect: temporal entrainment in overlapped speech," in *Proceedings of Speech Prosody 2012*, Accepted for publication.

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

[3] 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.

[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. 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.

[7] 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.

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

[9] Š. Beňuš, "Are we 'in sync': turn-taking in collaborative dialogues," in *Proceedings of Interspeech 2009*, Brighton, U.K., 2009, pp. 2167–2170.

[10] 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.

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

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

[13] 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.

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

[15] M. Włodarczak, J. Šimko, and P. Wagner, "Evidence for coordination between overlap onsets and inter-stress intervals," in *Perspectives on Rhythm and Timing Workshop*, Accepted.

[16] 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.

[17] R. Bertrand, P. Blache, R. Espesser, G. Ferré, C. Meunier, B. Priego-Valverde, and S. Rauzy, "Le CID — Corpus of Interactional Data — Annotation et exploitation multimodale de parole conversationnelle," *Traitement Automatique des Langues*, vol. 49, no. 3, pp. 1–30, 2008.

[18] K. J. Kohler, B. Peters, and M. Scheffers, "The Kiel corpus of spontaneous speech, vol. IV. German: Video Task Scenario (Kiel DVD No. 1)," http://www.ipds.uni-kiel.de/kjk/pub_exx/kk2006_6/InfoDVD1.pdf, 2006.

[19] J. Morton, S. Martin, and C. Frankish, "Perceptual centers (p-centers)," *Psychological Review*, vol. 83, pp. 405–408, 1976.

[20] G. Allen, "The location of rhythmic stress beats in english: An experimental study I & II," *Language and Speech*, vol. 15, pp. 72–100, 1972.

[21] K. Rapp, "A study of syllabic timing," *Speech Transmission Laboratory, Royal Instit. Technology Quarterly Status and Progress Report*, vol. 1, pp. 14–19, 1971.