INTERSPEECH 2013

Pitch and duration as a basis for entrainment of overlapped speech onsets

Marcin Włodarczak, Juraj Šimko, Petra Wagner

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

Abstract

The present paper reports on the impact of pitch accents and duration on temporal organisation of overlapping speech onsets in spontaneous dialogue. We observe a non-random pattern of overlap initiations within intervals between consecutive pitch accents, thus extending our earlier reports of a similar effect within vowel-to-vowel intervals. The latter finding was interpreted as a tendency to start overlapped speech directly before perceptually prominent vocalic onsets. In an attempt to reconcile these results, we investigate whether the effect observed on vowel-to-vowel intervals is influenced by presence of pitch accents and lengthening, both of which are known to be correlated with perceptual prominence. We find a strong effect of duration, which, however, does not on its own account fully for the observed pattern, indicating that other correlates of prominence might be involved in guiding the timing of overlap onsets.

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

1. Introduction

In dialogue, *when* something is said can be equally important to *what* is said. Timely initiation of speaking turns with avoidance of gaps and overlaps has long been put forward as a foundation of speaker exchange mechanism in conversation [1, 2]. The proposed precision of turn end prediction has led some authors [3, 4] to suggest that interlocutors are able to achieve the prescribed timing of their speech with respect to speech of their dialogue partner because they pick up on each other's speech rhythm and become *entrained* to one another.

However, more recently claims of the incredible accuracy of speaker exchange were called into question when frequencies of silences and overlaps were found to be higher than previously assumed [5]. Similarly, the rhythm-based models of turn-taking, backed by little empirical support even in their original formulation, failed to be reproduced in larger-scale corpus-based studies [6, 7, 8]

At the same time, rhythmicity has been found to underlie large areas of human social behaviour [9] and has also been noted for patterns of speech and silence in dialogue [10]. Relatedly, synchronisation between interlocutors has been demonstrated for behaviours such as body sway [11] and gazing [12]. In a similar vein, completely untrained speakers were found to be extremely adept at synchronising with each other when asked to read a piece of text in parallel [13]. These results have been interpreted in terms of coordination patterns linking phenomena that span speech perception and motor action.

To date very little effort has been directed to combine the research on phonetic convergence, turn-taking and temporal synchronisation among speakers. Our aim in this and earlier papers has been to bridge this gap by investigating temporal patterns in initiation of overlapped speech onsets, and thus provide the missing empirical evidence for temporal synchronization underlying turn-taking mechanisms.

Specifically, we investigate where overlap onsets occur relative to landmarks, such as syllabic boundaries, in the other speaker's turn. Previously we reported results indicating a non-random pattern of timing overlap onsets within syllables and intervals between consecutive vowel onsets (henceforth vowel-to-vowel intervals, VTV) [14, 15, 16]. Similar temporal patterns were observed in four analysed languages: English, French, German and Finnish. The results are reproduced in Figure 1, which plots distributions of normalised overlap onsets within the first overlapped VTV in interlocutor's speech. Normalised onset time on the abscissa depicts the proportion of the VTV interval elapsed before the overlap onset, i.e., the value of 0.5 corresponds to an overlap initiated at precisely mid-point between consecutive vowels onsets.

In all data sets overlap onsets were most likely to start around 80% of the VTV duration. As syllabic boundaries followed no consistent pattern across the languages, vowel onsets might provide a more robust guidance for inter-speaker entrainment, and the observed phenomenon might be more adequately described in terms of perceptually salient *p-centres* [17] rather than abstractly defined phonological events. This argument is further reinforced by the fact that the languages in question are conventionally classified as belonging to different rhythmic types (*syllable-timed* for French, *stress-timed* for English and German and *mora-timed* for Finnish), suggesting limited utility of these categories to explaining the observed effect and a common perceptual basis for speaker coupling.

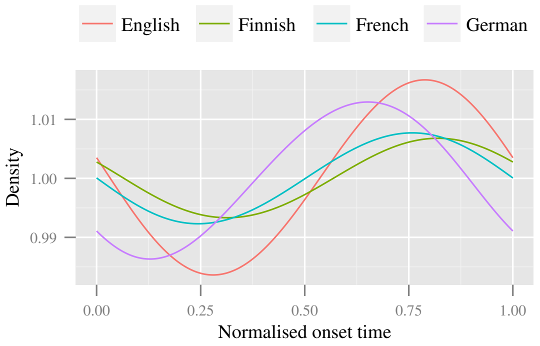

Figure 1: Distribution of overlap onsets normalised to the duration of the first overlapped VTV in interlocutor's turn (reproduced from [16]).

25 – 29 August 2013, Lyon, France

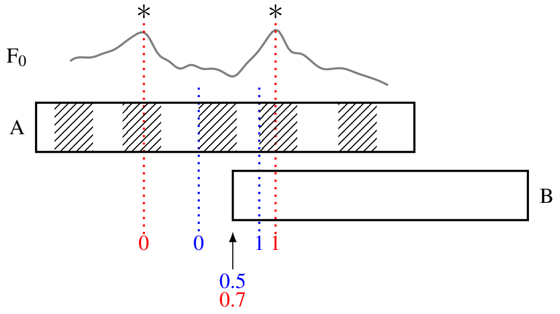

Figure 2: Overlap onset relative to the duration of the first coinciding inter-accent interval (red) and vowel-to-vowel interval (blue) in overlappee's speech. The stripes represent speakers' IPUs, vocalic intervals are marked with shaded areas, and pitch accents with stars.

In the present paper we pursue this line of inquiry by evaluating the influence of pitch and duration, both of which have been consistently shown to be linked to perceptual prominence in speech [18], on temporal patterning of overlapped speech onsets in English.

2. Method

We used a subset of the Switchboard corpus [19] annotated with pitch accent labels [20]. The data comprised 75 spontaneous telephone conversations on pre-defined topics between strangers. Stretches of overlapping speech were calculated from *MS-State* word-alignments [21] concatenated into inter-pausal units (IPUs) bounded by at least 100 ms of silence and/or laughter. For each overlap, the first overlapped interval between consecutive pitch accents (inter-accent interval, IAI) was identified. Overlap onsets were then normalised relative to the duration of this interval: the *IAI-normalised onset time* was calculated by dividing the duration of the interval between the previous pitch accent and the onset of the overlapping utterance by the duration of the overlapped IAI.

Additionally, to allow a direct comparison with our earlier results and to evaluate the effect of pitch accents on timing of overlaps within VTVs, the *VTV-normalised onset time* was calculated for this corpus subset in a similar fashion by dividing the duration of the interval between the previous vowel onset and the onset of the overlapping utterance by the duration of the first overlapped VTV. The procedure is illustrated schematically in Figure 2.

Overlaps coinciding with VTVs preceding the first and following the last pitch accent in overlappee's IPUs were excluded from the analysis with a view to eliminating simultanous starts and terminal overlaps, which are more likely to show effects of utterance boundary prediction than of continuous inter-speaker coordination. Overall 827 overlaps were analysed. No distinction was made between collaborative and competitive overlaps.

3. Results

The distribution of IAI-normalised overlap onset time is plotted in the left panel of Figure 3. A statistically significant result of Kuiper's test [1] comparing the observed distribution against a uniform distribution ($p < 0.01$) indicates non-randomness of overlap initiation. Specifically, overlaps are least likely to be initiated around pitch accents with a broad peak around the 60 % of the IAI duration, i.e. they are more frequent in the latter half of the IAI.

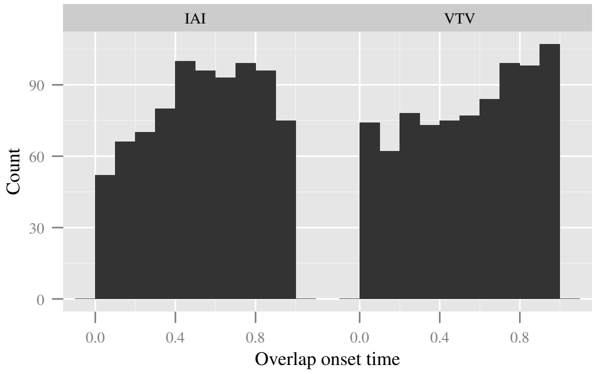

Figure 3: Distribution of overlap onset time normalised to the duration of IAI (left) and VTV (right) coinciding with the overlap onset.

For comparison, VTV-normalised onset time calculated on the same 75 Switchboard dialogues is presented in the right panel of Figure 3. The distribution is significantly different from random ($p < 0.01$) and similar in shape to distributions obtained for the whole corpus and the other languages analysed previously (see Figure 1).

Given the non-random timing of overlap initiation both within VTVs and IAIs, a question arises as to the relation between the two observed effects. Can one be explained in terms of the other? Can they be both traced down to a common cause? Or are they more or less independent of each other?

To answer these question and separate the various possible contributing factors, the VTV distribution in Figure 3 was split depending on whether the VTV coinciding with the overlap onset carried pitch accent. If presence of pitch accents was indeed the main cause of the observed effect, non-accented VTVs should follow a random pattern. The resulting distributions, plotted in Figure 4, are in line with this hypothesis: the distribution of accented VTVs is markedly less flat than that for non-accented ones, and unlike the latter is significantly different from a uniform distribution (p smaller than 0.001 and equal to 0.085 for accented and non-accented VTVs respectively).

While the above results suggest that presence of F_0 movement has an effect on timing of overlap onsets, lack of statistically significant outcome for unaccented VTVs certainly does not warrant inferring lack of effect. Additionally, the accented / non-accented dichotomy potentially conflates pitch and durational factors. With a view to separating individual contributions of pitch movement and duration, accented and non-accented VTVs were further split on their median durations. The resulting distributions are plotted in Figure 5.

Each of the categories in Figure 5 was compared against a uniform distribution yielding the following p-values: < 0.001 (long accented, long non-accented), 0.95 (short accented), 0.64 (short non-accented). The fact that the distribution of long non-

[1] Kuiper's test should be used as an alternative to Kolmogorov-Smirnov test when, as here, the quantities measured are of cyclic nature, that is when the location of 0 is purely arbitrary. The traditional Kolmogorov-Smirnov test is used below when data is split into categories yielding non-contiguous units, e.g. accented/non-accented VTVs.

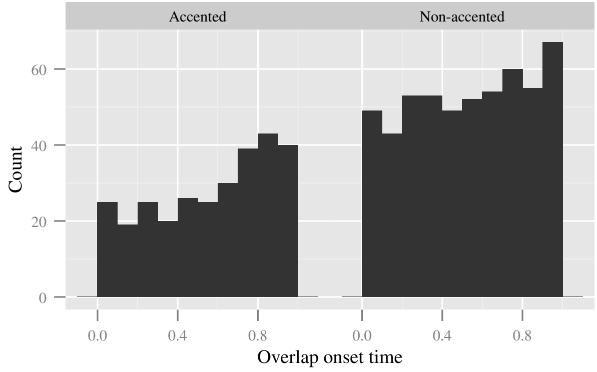

Figure 4: Distribution of normalised overlap onset time within accented and non-accented VTVs.

accented VTVs displays a non-random pattern both statistically and visually whereas that of short accented ones does not might indicate that duration is the main (or indeed the only) factor influencing overlap initiation patterns with little contribution of pitch. However, these results need to be taken with caution given the small counts, especially in the accented category, and the known subtlety of inter-speaker adaptation phenomena. Indeed, while duration might be the main perceptual cue driving inter-speaker entrainment, it could be mediated by pitch modulation. More generally, individual contributions of the two features will be difficult to disentangle because of the inherent lengthening effect of accentuation (in our data accented and non-accented VTVs have mean durations of 291 and 190 ms respectively).

These reservations are confirmed by evidence presented in Figure 6, in which overlap onset values were plotted separately for long and short VTVs (again split on median duration, equal to 220 ms) in the entire Switchboard corpus. Two things are of import here. First, short syllables exhibit a weak but non-random pattern ($p < 0.01$) similar to that reported for the other languages. However, the effect is very subtle and, therefore, is difficult to ascertain with small sample sizes. Second, in long VTVs the likelihood of overlaps tends to rise towards the end of the unit with no indication of a similar increase near the beginning. This is surprising since many of the overlaps produced in the vicinity of vowel onsets should be expected to fall into the next VTV, thereby producing the other half of the tail. It, therefore, suggests that the effect observed on long VTV is not merely quantitatively but also qualitatively different from that on short VTVs. In particular, long VTVs might invite overlaps by virtue of being a cue to disfluencies or indicating other production problems. Since these cases correspond to a one-off reaction to a stimulus, they do not constitute inter-speaker entrainment proper. However, it is not clear how they could be separated in the analysis above.

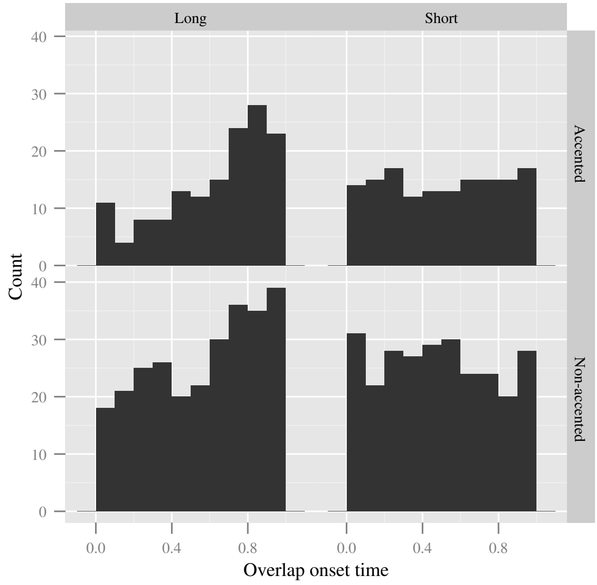

Figure 5: Distribution of normalised overlap onset time within accented and non-accented VTVs split on median durations.

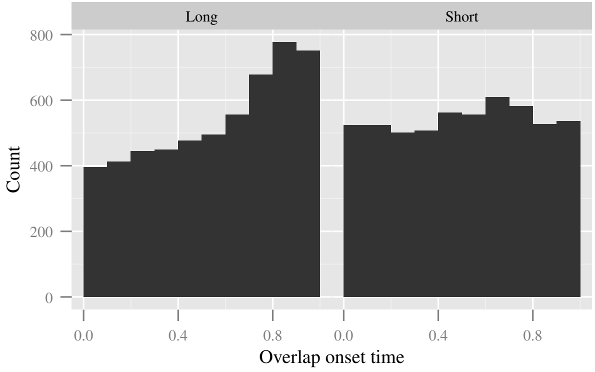

Figure 6: Distribution of normalised overlap onset time within short and long VTVs split on median durations.

4. Discussion and conclusions

This paper attempted to investigate contribution of perceptual prominence as a basis for inter-speaker entrainment. Overlap onsets were found to be distributed non-uniformly within intervals between consecutive pitch accents, attesting to presence of a pitch-related effect. Specifically, the shape of the distribution indicates a decreased likelihood of overlap onsets in the vicinity of pitch accents. Since prominence in English has been described as related primarily to pitch, this finding is in line the hypothesised role of perceptual prominence in guiding inter-speaker entrainment.

Subsequent analyses sought to relate the effect observed on IAIs to the previously reported VTV effects of an increased likelihood of overlap initiation directly before a vocalic onset. However, the obtained results are by no means straightforward. Although, accented VTVs were on the whole found to exhibit a stronger pattern than non-accented VTVs, results in Figure 5 suggest that the effect can be mainly attributed to duration rather than to presence of pitch accents. Generally, long VTVs exhibited a markedly non-random pattern, regardless of their accentedness. By contrast, short VTVs, whether accented or not, were not found to deviate significantly from a random baseline. Thus, there is little evidence for the impact of accentedness on timing of overlaps within VTVs. Insofar as the hypothesised link between temporal patterns in overlap onsets and prominence is correct, this is consistent with those accounts of prominence in English which have described it mainly in terms of durational features [22].

It should be borne in mind, however, that the number of data points in the analysed categories might be too small to allow detecting subtler influences of vocalic onsets and accentuation. Indeed, the analysis of overlap onsets within short VTVs in the entire corpus revealed a significant albeit weak effect within those segments too. This effect is likely to be easily overridden by the strong influence of duration.

Moreover, the effect obtained on short VTVs was qualitatively different from the effect on long VTVs, implying that duration alone cannot account fully for the observed pattern. Indeed, it appears plausible that at least some of the durational effects correspond to overlaps produced in response to lengthening in overlappee's speech, for example indicative of production problems, and might in fact obscure the ongoing temporal coordination between speakers. Nevertheless, for lack of a statically significant result within short non-accented VTVs in the portion of the corpus labelled for pitch accents, a definite conclusion concerning the relationship between the influences of pitch accents and vocalic onsets, and the likely interaction of duration, is not possible at present.

A task related to ours was pursued by Cummins, who in a series of experiments with modified stimuli [13] attempted to assess the contribution of various acoustic features to successful synchronisation of a text read in parallel by two speakers. Although people have been observed to be extremely skilled at this task, little is known about the properties of the signal which allow such tight inter-speaker coupling. Cummins' results, however, were far from unequivocal. The hypothesised importance of F_0 contours and of amplitude envelope were only partly borne out, since neither of the features on its own provides sufficiently strong cues for synchronisation. Consequently, the results point towards an intricate interplay of all the factors.

It seems likely that a similar interplay between various features might be at work in speaker synchronisation in dialogue. In this paper we have demonstrated that timing of overlap onsets is influenced by presence of pitch accents. However, disentangling the many links between pitch, duration and segmental prominence remains an enterprise of the future, and one whose results will possibly depend on the language under investigation.

5. Acknowledgements

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

6. References

[1] 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.

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

[3] 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.

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

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

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

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

[8] M. O'Dell, M. Lennes, and T. Nieminen, "Modeling turn-taking rhythms with oscillators," *Linguistica Uralica*, vol. 3, pp. 218–227, 2012.

[9] E. D. Chapple, *Culture and biological man: Explorations in behavioral anthropology*. New York: Holt, Rinehart and Winston, 1970.

[10] R. M. Warner, "Periodic rhythms in conversational speech," *Language and Speech*, vol. 22, no. 4, pp. 381–396, 1979.

[11] K. Shockley, A. A. Baker, M. J. Richardson, and C. A. Fowler, "Articulatory constraints on interpersonal postural coordination." *Journal of Experimental Psychology: Human Perception and Performance*, vol. 33, no. 1, pp. 201–208, 2007.

[12] D. C. Richardson, R. Dale, and J. M. Tomlinson, "Conversation, gaze coordination, and beliefs about visual context," *Cognitive Science*, vol. 33, no. 8, pp. 1468–1482, 2009.

[13] F. Cummins, "Rhythm as entrainment: The case of synchronous speech," *Journal of Phonetics*, vol. 37, no. 1, pp. 16–28, 2009.

[14] M. Włodarczak, J. Šimko, and P. Wagner, "Syllable boundary effect: temporal entrainment in overlapped speech," in *Proceedings of Speech Prosody 2012*, 2012, pp. 611–614.

[15] ——, "Temporal entrainment in overlapped speech: Cross-linguistic study," in *Proceedings of Interspeech 2012*, Portland, OR, 2012.

[16] M. Włodarczak, J. Šimko, P. Wagner, M. O'Dell, M. Lennes, and T. Nieminen, "Finnish rhythmic structure and entrainment in overlapped speech," in *Nordic Prosody. Proceedings of the XIth Conference*, E. L. Asu and P. Lippus, Eds. Frankfurt am Mein: Peter Lang, 2013, pp. 421–430.

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

[18] D. B. Fry, "Experiments in the perception of stress," *Language and speech*, vol. 1, no. 2, pp. 126–152, 1958.

[19] 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.

[20] 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.

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

[22] R. Silipo and S. Greenberg, "Automatic transcription of prosodic stress for spontaneous English discourse," in *Proceedings of the XIVth International Congress of Phonetic Sciences*, vol. 3, 1999, pp. 2351–2314.