Context cues for classification of competitive and collaborative overlaps

Catharine Oertel[1,2]*, Marcin Włodarczak*[3]*, Alexey Tarasov*[4]*, Nick Campbell*[1]*, Petra Wagner*[3]

[1]Speech Communication Lab, Trinity College Dublin, Ireland
[2]KTH Speech, Music and Hearing, Stockholm, Sweden
[3]Faculty of Linguistics and Literary Studies, Bielefeld University, Germany
[4]Digital Media Centre, Dublin Institute of Technology, Ireland

oertelgc@tcd.ie, mwlodarczak@uni-bielefeld.de, aleksejs.tarasovs@student.dit.ie,
nick@tcd.ie, petra.wagner@uni-bielefeld.de

Abstract

Being able to respond appropriately to users' overlaps should be seen as one of the core competencies of incremental dialogue systems. At the same time identifying whether an interlocutor wants to support or grab the turn is a task which comes naturally to humans, but has not yet been implemented in such systems. Motivated by this we first investigate whether prosodic characteristics of speech in the vicinity of overlaps are significantly different from prosodic characteristics in the vicinity of non-overlapping speech. We then test the suitability of different context sizes, both preceding and following but excluding features of the overlap, for the automatic classification of collaborative and competitive overlaps. We also test whether the fusion of preceding and succeeding contexts improves the classification. Preliminary results indicate that the optimal context for classification of overlap lies at 0.2 seconds preceding the overlap and up to 0.3 seconds following it. We demonstrate that we are able to classify collaborative and competitive overlap with a median accuracy of 63%.

Index Terms: overlapping speech, dialogue prosody, cooperative overlaps, competitive overlaps

1. Introduction

The assumption that dialogue participants try to minimise gaps and overlaps between successive turns [1] has been taken to be accurate for a long time. Only in the last few years has it been shown that overlap is a wide-spread phenomenon in dialogue. Quantitative studies have found about 40% of all between-speaker intervals to be overlaps [2]. Therefore, the no-gaps-no-overlaps assumption does not seem to hold to the degree previously assumed.

A widely adopted categorisation of overlaps is one into competitive overlaps, in which the incoming speaker attempts to forcefully take over the turn, and collaborative ones, in which the incoming speaker assists the current speaker in his or her speech. Phonetic characteristics of both types of overlaps were first studied by French and Local [3], who found that interlocutors consistently mark their utterances as competitive by using increased pitch and loudness alone, regardless of the semantic and syntactic features.

Wells and Macfarlane [4] investigated competitiveness of overlaps in relation to their position within the ongoing turn. While reproducing French and Local's finding about the relevance of pitch and loudness for turn competitiveness, they found that overlapping speech is treated by dialogue participants as turn competitive only if it is initiated before a transition relevance place (TRP). By contrast, overlaps starting at TRPs are not treated as turn competitive. Additionally, the authors identified TRP-projecting accents, which are phonetically distinct from non-TRP-projecting ones. In that light, a competitive overlap could be characterised as preceding a TRP-projecting accent and produced with increased pitch and loudness. The relevance of pitch and intensity has been also further corroborated by Kurtic, Brown and Wells [5] and Yang and Heeman [6]. Kurtic, Brown and Wells [7] carried out classification experiments on overlapping talk distinguishing between competitive and non-competitive overlaps. They found duration to be the most distinguishing feature. Qualitative research into cooperating and intrusive overlaps has revealed both cultural [8] and gender [9] differences.

Importantly, most previous studies have only considered characteristics of overlapping speech itself. Indeed, French and Local claim that the phonetic features characterising overlapping speech are rarely carried on beyond the end of an overlap. However, in real-life contexts classification of overlapping speech is difficult. For example, problems with cross-talk might occur, which make extraction of meaningful F0 or intensity measurements difficult. If information that an overlap occurs was also signalled in the surrounding speech, it would facilitate building systems capable of much more sophisticated turn-taking behaviour. Current state of the art incremental dialogue systems such as [10] would be well suited to include information on overlap type. Such dialogue systems would be perceived as more natural and human-like if they were capable of intelligent overlap management.

Therefore, in this paper we present the first results on prosodic and body movement features of the speech signal surrounding overlaps. Firstly, we study prosodic features calculated over a large time window (5 seconds). We hypothesise that intensity, median F0 and F0 range will be higher in windows containing overlapping speech.

Secondly, we examine prosodic and body movement cues in the direct vicinity of overlaps for the classification of collaborative and competitive classes of overlap. In particular, we examine how far these cues extend beyond the onset and offset of an overlap by extracting features from increasingly large time windows. We hypothesise that cues extracted in closer vicinity of an overlap should produce better classification results.

The first two authors are listed in alphabetical order.

2. Data

All analyses in this study are carried out on the D64 corpus [11]. The D64 corpus is a multi-modal corpus recorded over two successive days resulting in approximately 8 hours of video recordings. The conversation was in no way directed or task-oriented. Five people participated in the recordings, 3 male and 2 female. For this study, we selected two 30 minute sections. The first chosen section contained social chat whereas the second section was mainly filled with technical talk.

3. Methodology

Discussed below are the annotation, feature extraction and classification procedures.

3.1. Annotation of Overlap

Overlap was annotated in the phonetic software "Praat" [12]. The start and end point of overlapping speech as well as the identities of the speakers involved in an overlap were annotated. Given that overlaps were drawn from a multi-party conversation there were instances in which more than 2 people overlapped. Similar to Kurtic et al [7] these instances were excluded from the analysis. For our analysis we use a set of 143 overlaps.

3.2. Annotation of Classes of Overlap

We distinguished between competitive and collaborative overlaps. We define competitive overlap as "One speaker is competing with another speaker for the right to continue speaking. The overlapping speaker may want to change the topic of the conversation or voice his own opinion on the topic discussed." We define collaborative overlap as "The overlapping speakers intention is not to interrupt the current speaker but rather assist the speaker in what he or she is saying." Also included in the collaborative category were backchannels, defined as "short affirmative signals to communicate that the speaker has heard and/or understood what the other person said". Overlap was annotated by two expert raters. All differences were discussed and manually resolved. Overall, 92 overlaps were collaborative and 51 were competitive.

3.3. Feature Extraction

In order to investigate the prosodic makeup in the vinity of overlapping speech, we extracted the following features from non-overlapping 5 second windows: median F0, F0 range and mean intensity. To normalise for speaker differences each speaker's feature values were converted to z-scores. If more than one participant spoke in a window, a mean value weighted by durations of individual utterances was calculated. Therefore, the time each participant spent speaking was taken into account.

In order to evaluate how far cues useful for the classification of collaborative and competitive overlaps extend beyond overlap boundaries we performed a series of classification tasks using features extracted from increasingly large preceding and following contexts from 0.2 s[1] up to 1 s with 0.1 s time step (see Figure 1).

Informed by results of related studies [5, 6], which found F0 and intensity patterns relevant for overlap resolution, we used the following features: minimum, maximum, mean and median F0, F0 range and mean intensity.

[1]Contexts shorter than 0.2 s were excluded from the analysis since they produced too many missing values.

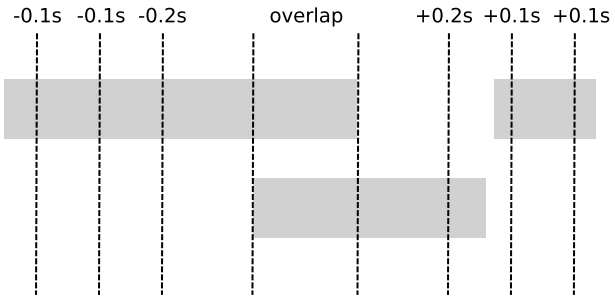

Figure 1: Feature extraction procedure. The grey stripes represent speakers' turns.

Additionally, averaged body and face movement of two participants calculated on a frame-by-frame basis was used. Coordinates of the body and the face (given by the location of its top left and bottom right corners) in each frame were extracted from the video data following [13] using the standard Viola-Jones algorithm [14]. As these coordinates are highly dependent on the distance of the person to the camera normalisation is carried out in order to obtain relative movement over the size of the detected face and body. A moving average was calculated only if the face was recognised.

For each overlap we also extracted the number and the total duration of overlaps preceding it by 1, 5, 30 and 60 s. This way we kept track of overlaps in dialogue history. The same normalisation methods as those discussed above were used.

3.4. Classification

We used support vector machines (SVMs) and compared their accuracies on varying context sizes preceding and following an overlap. For optimisation purposes we carried out parameter selection. Three parameters have been tuned: type of kernel (linear or RBF), cost parameter C (for both kernels) and γ parameter of the RBF kernel.

Due to the small number of training samples, we used a relatively small number of folds $N = M = 3$ and the following cross-validation method: Let the number of folds for cross-validation be N. In order to tune parameters, data points from $N - 1$ folds of training data are divided into M folds, then for each combination of parameters a separate classifier is trained on $M - 1$ folds and tested on the remaining fold. Then the classifier with the highest accuracy is selected and trained on $N - 1$ "inner" folds and tested on the remaining one. The same process is repeated for each of N folds. An averaged performance score on all N folds is reported as a performance on the whole dataset. Thus, the parameters were tuned separately for each classifier.

In order to counteract overfitting on the small training set the experiment was carried out 100 times with different distributions of instances across N folds. Median values across 100 runs are reported as the classification accuracy. By doing so we ensured that the reported performance scores show the generalisation ability of the classifiers on a bigger dataset.

Friedman test was used to test the statistical significance of the differences in accuracies. Averaged class accuracy is used as the measure of classifier performance. Therefore, our baseline for classification at chance equals 50%.

We carried out experiments which used the preceding context alone, the succeeding context alone, the combination of the two for all time intervals (from 0.2 to 1.0 with 0.1 increments).

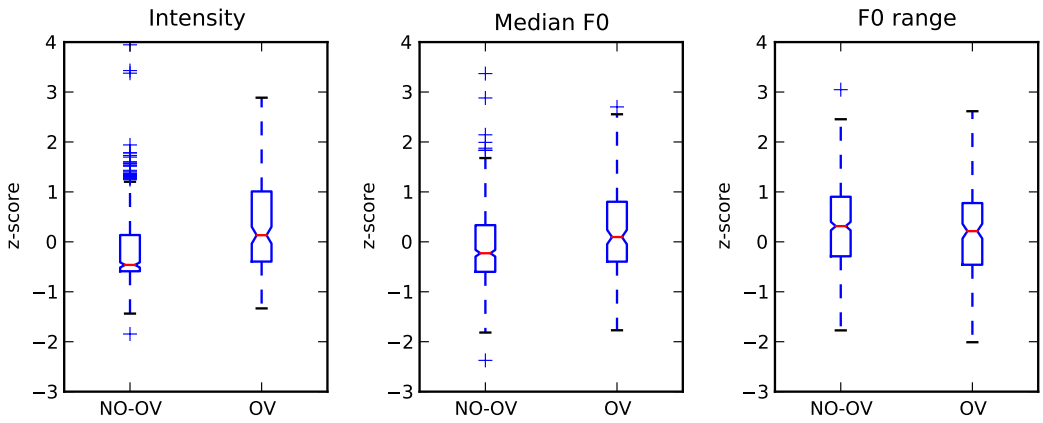

Figure 2: Distributions of mean intensity, median F0 and F0 range in 5 second windows containing non-overlapping speech only (NO-OV) and containing overlapping speech (OV).

4. Results

The first question we addressed is whether the sound signal surrounding overlaps is different from the sound signal in proximity of non-overlapping speech. Figure 2 shows that this is indeed the case. Median F0 (Mann-Whitney U, $p < 0.001$), intensity (Mann-Whitney U, $p < 0.001$) as well as F0 range (Mann Whitney U, $p < 0.05$) are significantly different in 5 s windows containing overlaps than in windows without them. Specifically, speech in proximity of overlaps is characterised by higher fundamental frequency and intensity as well as smaller fundamental frequency range.

The second question was whether prosodic and body movement features from preceding and following contexts can be used to correctly classify instances of overlap as collaborative or competitive and if so how far these cues extend beyond overlap boundaries. Figure 3 shows distributions of averaged class accuracies for each preceding and following context. As can be observed, median accuracies for all contexts fall around 0.5, which is the accuracy expected by chance for a two-class problem. The only contexts sizes which reached higher accuracies were 0.2 s for the preceding context and 0.2 and 0.3 s for the following context.

Mann-Whitney U test was used to compare the prosodic features of collaborative and competetive 200 ms contexts. While none of the differences were significant for the preceding context, minimum, maximum, mean and median F0 as well as mean intensity of the speech following competetive overlaps were significantly higher in the case of collaborative overlaps.

Additionally, Friedman test was used to compare the right and left contexts but the difference was not statistically significant. In other words, neither side significantly outperforms the other one.

Plotted in Figure 4 are distributions of accuracies for features from the right and left contexts together. Here the 0.2 and 0.3 s contexts again outperform the remaining ones. Moreover, the accuracies for longer context and the overall decline of accuracy is less abrupt than in the case of either preceding or following contexts alone.

5. Discussion

We found that speech surrounding overlaps has a characteristic prosodic profile. Importantly, the differences in fundamental frequency and intensity were significant for features averaged over windows as long as 5 seconds. Windows of that size are likely to contain multiple turns from more than one speaker. Therefore, one plausible explanation for the observed differences is that speech in proximity of overlaps reflects increased involvement in the conversation on the part of interlocutors [15].

The classification results indicate that prosodic and body movement information from the preceding and the following contexts can be used for classification of collaborative and competitive overlaps. However, more important than absolute values are the relative accuracies for different context sizes. They indicate that cues can be found only in the direct vicinity of an overlap and that they do not exceed 0.2 s for the preceding context and 0.3 s for the following context. These contexts sizes also outperform longer contexts when features from the right and the left contexts are used fused. This is in line with French and Local's [3] claim that features characterising competitive overlaps are sometimes carried over the end of an overlap but never by more than a foot. At the same time, the authors seem to claim that these features are never found before the overlap onset, which is not the case in our data. However, the fact that the impact of overlaps on their immediate context is sometimes limited, might partly explain the variance of accuracies depending on the combination of folds if such instances are not distributed uniformly between the training and test sets.

6. Conclusions

We presented two main findings concerning the characterstics of prosodic and body movement features in the vicinity of overlapping speech. On the one hand, the speech signal around overlaps is prosodically different from the speech signal in the vicinity of non-overlapping speech. This effect seems to extend far beyond overlap boundaries and could be attributed to speakers' increased involvement in the conversation. On the other hand, 200-300 ms directly preceding and following the overlap can be potentially useful for classification of overlaps into the collaborative and competitive categories. It could be also treated as first step towards *prediction* of these overlap classes.

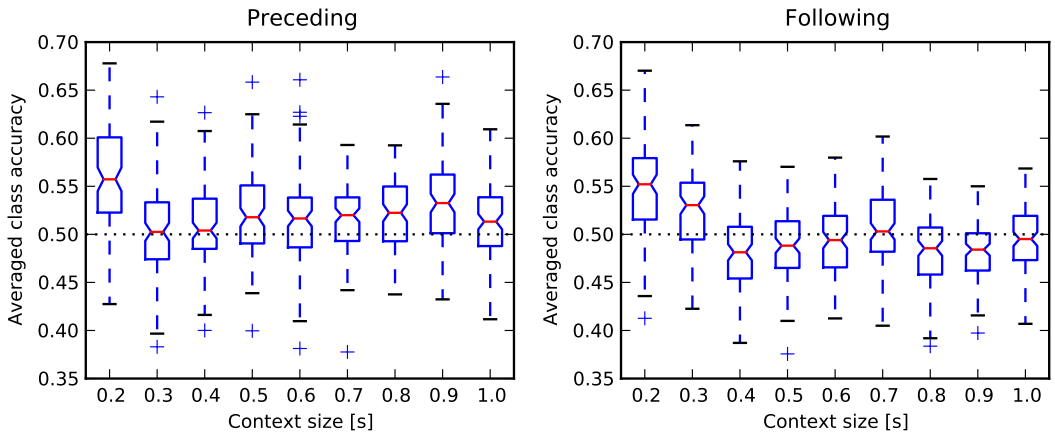

Figure 3: Classification accuracies for preceding and following contexts used separately.

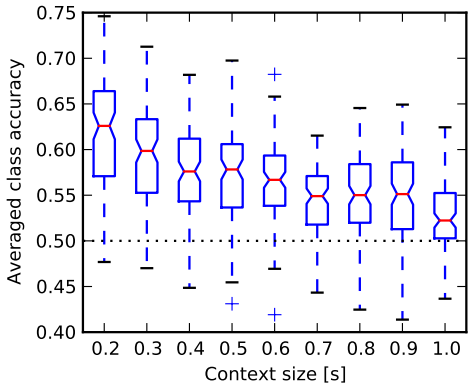

Figure 4: Classification accuracies for joined contexts.

While preliminary, our results can be useful for designing attentive spoken dialogue systems. Being able to respond appropriately to users' overlaps should be seen as one of the core competencies of such systems. The results also indicate that overlap frequency could be used to monitor user's involvement in communication.

7. Acknowledgements

Catharine Oertel is supported by the Irish Research Council for Science Engeneering and Technology–Embark Initiative. Marcin Włodarczak is supported by the German BMBF-funded "Professorinnenprogramm" FKZ 01FP09105A. Alexey Tarasov is supported by the Science Foundation Ireland under Grant No. 09-RFP-CMS253. The work was supported in part by the Riksbankens Jubileumsfond (RJ) project P09-0064:1-E Prosody in conversation and the Swedish Research Council (VR) project 2009-1766 The rhythm of conversation.

We would also like to thank Daniel Neiberg, Jens Edlund and Zofia Malisz for valuable comments.

8. References

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

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

[3] P. French and J. Local, "Turn-competitive incomings," *Journal of Pragmatics*, vol. 7, pp. 17–38, 1983.

[4] B. Wells and S. Macfarlane, "Prosody as an interactional resource: Turn-projection and overlap," *Language and Speech*, vol. 41, no. 3/4, pp. 265–294, 1998.

[5] E. Kurtic, G. J. Brown, and B. Wells, "Fundamental frequency height as a resource for the management of overlap in talk-in-interaction," in *Where Prosody Meets Pragmatics*, D. Barth-Weingarten, N. Dehe, and A. Wichmann, Eds. Bingley: Emerald, 2009, pp. 183–204.

[6] F. Yang and P. A. Heeman, "Initiative conflicts in task-oriented dialogue," *Computer Speech and Language*, vol. 24, no. 2, pp. 175–189, 2010.

[7] E. Kurtić, G. J. Brown, and B. Wells, "Resources for Turn Competition in Overlap in Multi-Party Conversations: Speech Rate, Pausing and Duration," in *Interspeech 2010*, Makuhari, 2010, pp. 2550–2553.

[8] K. Murata, "Intrusive or co-operative? A cross-cultural study of interruption," *Journal of Pragmatics*, vol. 21, pp. 385–400, 1994.

[9] M. Makri-Tsilipakou, "Interruption revisited: Affiliative vs. disaffiliative intervention," *Journal of Pragmatics*, vol. 21, pp. 401–426, 1994.

[10] D. Schlangen and G. Skantze, "A General, Abstract Model of Incremental Dialogue Processing," *Dialogue & Discourse*, vol. 2, no. 1, pp. 83–111, 2011.

[11] C. Oertel, F. Cummins, N. Campbell, J. Edlund, and P. Wagner, "D64: A corpus of richly recorded conversational interaction," in *Proceedings of LREC 2010; Workshop on multimodal corpora:advances in capturing, coding and analyzing multimodality*, Valetta, 2010, pp. 27–30.

[12] P. Boersma and D. Weenink, "Praat: doing phonetics by computer." [Online]. Available: http://www.praat.org

[13] S. Scherer, N. Campbell, and G. Palm, "Multimodal laughter detection in natural discourses." in *Proceedings of the 3rd international workshop on human-centered robotic systems (HCRS'09)*, 2009, pp. 111–121.

[14] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[15] C. Oertel, S. Scherer, and N. Campbell, "On the use of multimodal cues for the prediction of involvement in spontaneous conversation," in *Interspeech 2011*, 2011, pp. 1541–1544.