Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee

Khiet P. Truong

Human Media Interaction, University of Twente
Enschede, The Netherlands
k.p.truong@utwente.nl

Abstract

One of the major properties of overlapping speech is that it can be perceived as competitive or cooperative. For the development of real-time spoken dialog systems and the analysis of affective and social human behavior in conversations, it is important to (automatically) distinguish between these two types of overlap. We investigate acoustic characteristics of cooperative and competitive overlaps with the aim to develop automatic classifiers for the classification of overlaps. In addition to acoustic features, we also use information from gaze and head movement annotations. Contexts preceding and during the overlap are taken into account, as well as the behaviors of both the overlapper and the overlappee. We compare various feature sets in classification experiments that are performed on the AMI corpus. The best performances obtained lie around 27%–30% EER.

Index Terms: overlapping speech, interruption, cooperative, competitive, classification

1. Introduction

Simultaneous speech (i.e. overlapping speech) is a relatively common phenomenon in spontaneous conversation. People may talk at the same time for various reasons. Usually, a distinction is made between overlaps that are cooperative or competitive of nature. Examples of cooperative overlaps are the expression of supportive agreement, or completion of an anticipated point [1, 2]. Cooperative overlaps are more supportive of the main speaker's floor rights, and the intention is to keep the attention on the main speaker's point. Competitive overlaps are disruptive and pose threats to the current speaker's territory by disrupting the process and/or content of the ongoing conversation [3, 4]. Competitive overlaps are typically made out of an urge of the speaker to attract the attention away from the ongoing speech [1]. This urge may be given in by the need to express something that is emotionally significant to the speaker. Several examples of competitive overlaps include the demand for new information or clarification, the expression of strong opinions or disagreement, shifting topic, or the intent to steal the floor [1, 2].

The distinction between competitive and cooperative overlaps can be important in social signal processing. Competitive overlaps are often seen as indicators of power, control, and dominance while cooperative overlaps can be associated with rapport displays.

Hence, in this paper, we investigate automatic ways of distinguishing between competitive and cooperative overlaps. We analyse how features extracted from various contexts surrounding the overlap, and extracted from the overlapper or overlappee influence classification performance. In addition to F0 and intensity, we use voice quality features since the quality of effort may help to identify competitive overlaps. Finally, focus of attention (i.e. gaze) and head movement information are also added as features.

In Section 2, we give an overview of related work. Section 3 describes the data used. Features and methods used are described in Section 4. Finally, we present our results and conclusions in Sections 5 and 6 respectively.

2. Related work

Several studies have shown that prosody plays an important role in analysing overlapping speech and the vicinity of overlapping speech. Although in some studies it is not directly clear whether a distinction is made between competitive and cooperative overlap, and whether what they call an interruption is similar to 'our' definition of competitive overlap, most studies indicate F0 and intensity as two major features in the analysis of overlaps.

In Shriberg et al. [5], it was studied whether overlapping speech could be predicted based on prosodic features. They were interested in finding out whether there is any correlation between the onset of overlaps and prosodic features of both the overlappee ('jump-in-points') and the overlapper ('jump-in-words'). Using decision trees, they achieved an accuracy of 64% for the task of classifying each word boundary as to whether or not any other speaker jumped in at that location. The results suggested that overlappers do not jump in at randomly chosen points in the speech but that there are certain contexts that are more likely to be jumped into. The features used in the tree indicate that speakers jump in at those locations that look similar to sentence boundaries but which are not actually sentence boundaries. For the 'jump-in-words' task, an accuracy of 77.8% was achieved. RMS and F0 were heavily used by the classifier and suggested that speakers raise their energy and voice when they interrupt the foreground speaker.

Other support for the observation that F0 and intensity are the main features in overlap classification comes from French and Local [6] (among others). French and Local [6] argue that it is the phonetic design of the incoming turn rather than its precise location that constitutes the overlap as turn-competitive. They argue that a competitive interruption is <h + f>, i.e., raised in pitch and loudness. This hypothesis was tested and supported by Wells and Macfarlane [7]. Support for this hypothesis was also provided by Yang [15] who investigated the prosodics of competitive and cooperative interruptions. There, it was concluded that competitive interruptions are typically high in pitch and amplitude due to the urge of the speaker to attract the attention away from the ongoing speech.

Recently, researchers have concentrated on other features then F0 and intensity alone, and in addition have looked at speech rate, disfluencies, body, hand, facial, and head movements for either the prediction of interruptions or the classification of competitive/cooperative overlaps. Oertel et al. [8] used prosodic features (from the overlapper) and body movment features (from both overlapper and overlappee) to investigate the context surrounding overlaps. Multimodal cues such as speech intensity, hand motions, and disfluencies were used in Lee et al. [9] to classify overlaps as either competitive or cooperative. Rather than classifying overlaps, Lee and Narayanan [10] aim to predict interruptions. Using the interruptee's acoustic features and the interrupter's facial and head movements, they come to the same conclusion as [5], namely that interruptions are not made at random points in the speech but are made in certain contexts that can be predicted. A similar conclusion was drawn by Gravano and Hirschberg [11]. Based on prosodic features, they find that interruptions do not occur at random locations but they are more likely to occur after certain types of IPUs (intonational phrase units). In addition, the onsets of interruptions yield significant differences in intensity, pitch level, and speech rate in comparison to other turn types. Speech rate was also investigated by Kurtic et al. [12] who, in contrast, did not find evidence that overlappers make use of fast speech rate to design the beginnings of their incomings as turn-competitive.

In summary, F0 and intensity are the dominant speech features used in overlap analysis, while body, facial, hand, and head movements are also used in recent studies. The contexts in which interruptions appear are not random contexts but have certain properties that can be predicted.

In contrast to previous studies, in the current study, we extract features from various contexts preceding the overlap or during the overlap, and compare how these perform in cooperative vs. competitive overlap classification. In addition to context, we also investigate to what extent features extracted from the overlappee are helpful. We introduce the use of voice quality features in addition to F0 and intensity to discriminate between competitive and cooperative overlaps. We add information about the focus of attention (i.e. gaze) and head movements to our set of acoustic features in order to see whether performance can be improved.

3. Data

We used data from the multiparty AMI Meeting corpus [13]. Five meetings[1] that contain multimodal annotations were selected for analysis. From these meetings, we automatically extracted overlap regions based on the dialog act annotations provided with the corpus. We listened to all of these overlap regions and excluded those that were not perceived as overlapping and those that are clear cases of backchannels. An additional criterion was that the overlapper and overlappee should be clearly identifiable which excludes complex overlap situations where more than two speakers speak at the same time. This selection procedure resulted in a number of 509 overlap instances that were annotated by 3 different annotators where each overlap instance has been labelled by 2 annotators. The annotators were asked to label each overlap instance as competitive (i.e., intrusive) or cooperative (i.e., non-intrusive, neutral) according to the following descriptions:

Competitive The overlapper disrupts the speech (breaks the flow) of the overlappee to take the turn and say something. The overlappee could be offended because he/she was not able to finish his/her sentence. Although the overlappee does not need to show that he/she is offended, the overlap could have been perceived as intrusive and/or competitive by the overlappee. The need to say something arises from the overlapper's own wants.

Cooperative The intention of the overlapper is to maintain the flow of the conversation, to coordinate the process and/or content of the ongoing conversation, and to offer help to the speaker when needed. The overlap does not abruptly disrupt the speech flow of the overlappee. It is most likely that the overlappee does not perceive this overlap as intrusive.

The average agreement found between the annotators amounts to Krippendorff's $\alpha = 0.30$. For the classification experiments only those overlap instances were used that were agreed upon by 2 annotators. This resulted in a final set of 355 overlap instances of which 140 are competitive (COMP) and 215 are cooperative (COOP).

4. Features and method

4.1. Features

For automatic extraction of the acoustic features, we use Praat [14]. For the head movement and focus of attention features, we use the manual annotations as provided with the AMI corpus.

4.1.1. Acoustic features

Previous literature points out that F0 and intensity are the two most distinctive features between cooperative and competitive overlaps. Hence, we extract F0 and intensity using a step time of 0.01s and an analysis window of 0.04s and 0.032s respectively. Our voice quality features are based on the Long-Term Averaged Spectrum (LTAS) that is extracted each 0.01s with an analysis window of 0.5s. Based on [15], we extract information about the distribution of energy in the LTAS in various frequency bands which is said to correlate with perceptual judgements of effort, breathiness, coarseness and head-chest register. More specifically, the maximum energy in the LTAS within a certain range of frequency range is used as a feature (for example, max(LTAS$_{\text{2k-5k}}$) refers to the maximum energy between the region of 2kH and 5kHz measured in the LTAS). In addition, the slope of the LTAS is used. Table 1 describes the acoustic features used in our study. Finally, we take the mean, standard deviation, maximum and minimum of these 0.01s frame-wise features over various speech intervals with certain lengths as defined in section 4.2. The total number of acoustic features used amounts to 32 (ACOUST). All features were speaker normalized by converting the feature values to z-scores where μ and σ are taken over the speech segments.

F0	
INTENS	
BREATHY1	[max(LTAS$_{\text{0-2k}}$) – max(LTAS$_{\text{2k-5k}}$)] – [max(LTAS$_{\text{2k-5k}}$) – max(LTAS$_{\text{5k-8k}}$)]
BREATHY2	max(LTAS$_{\text{2k-5k}}$) – max(LTAS$_{\text{5k-8k}}$)
EFFORT	max(LTAS$_{\text{2k-5k}}$)
COARSE	max(LTAS$_{\text{0-2k}}$) – max(LTAS$_{\text{2k-5k}}$)
HEADCHEST	max(LTAS$_{\text{0-2k}}$) – max(LTAS$_{\text{5k-8k}}$)
SLOPE_LTAS	slope of the LTAS

Table 1: *Acoustic features used for classification experiments.*

[1] IS1003b, IS1003d, IS1008b, IS1008c, and IS1008d

4.1.2. Head movements

For head movement features, we used the annotations provided with the AMI corpus. The following communicative head movement events were coded and used in our classification experiments: *concord* (signals agreement), *discord* (signals uncertainty or disagreement), *negative* (negative response), *turn* (effort by listener to take the floor), *deixis* (pointing gesture involving the head), *emphasis* (effort to highlight a word/phrase), and *other* (all other communicative head gestures). Additionally, it was marked whether each head gesture was expressed in the form of a shake or nod. Binary values were computed for the absence or presence of these events which resulted in an 9-dimensional feature vector (HEAD).

4.1.3. Focus of attention

For focus of attention (gaze) we used the 'foa' annotations provided with the AMI corpus. From these annotations, binary values were computed for when the overlapper or overlappee looked at another specified object (whiteboard, etc.), another unspecified object (everything else in the room) or another person. If the overlapper and overlappee looked at each other, we would label this as mutual gaze. We register the duration of mutual gaze. This results in an 4-dimensional feature vector (FOA, 3-dimensional when the features are extracted in the overlappee).

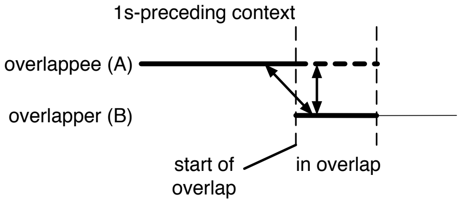


Figure 1: *Contexts for features. Arrows indicate differences between preceding context and 'in overlap'.*

4.2. Context, overlapper, overlappee

The features are extracted from various speaker regions in the speech signal. First, we extract features from the overlapper *during* overlap for various durations (i.e., 0.2s, 0.4s, 0.6s, 0.8s, and 1.0s). The set of acoustic features extracted for the overlapper while being in overlap is referred to as ACOUST_B_IN. Secondly, we take the acoustic features of the overlappee in the preceding context (i.e., 1s prior to the start of the overlap) into account by calculating the difference between ACOUST_B_IN and the overlappee's preceding context. This set of acoustic features is referred to as DIFF_AB. When the features of the overlappee and overlapper are both measured during overlap, we refer to this feature set as DIFF_AB_IN. Thirdly, the focus of attention and head movement features are measured for the overlappee in the 1s-preceding context which we refer to FOA_A and HEAD_A. When these features are measured for both the overlappee and overlapper during overlap, we refer to them as FOA_AB_IN and HEAD_AB_IN. In short, the affixes A and B refer to the overlappee and overlapper respectively, while IN refers to 'while being in overlap'.

4.3. Set-up classification experiments

Since the amount of data is relatively small, we performed a ten-fold cross validation. For each fold, the data was divided into 3 distinct data sets with a 80%-10%-10% split for training, developing, and testing respectively while maintaining the proportional distributions of the two classes. The 10% developing set was randomly sampled for each fold from the remaining 90% when the test set was excluded.

A Support Vector Machine (SVM) [16] with a Gaussian kernel was used as classification technique. The cost and gamma parameters were found on the dev set. As performance measures, Equal Error Rates (EERs) are reported.

5. Results

5.1. Classification experiments

As our focus is on the acoustic behavior of the overlapper we first experimented with the ACOUST_B_IN set which we consider our base set. For various lengths after the start of an overlap, acoustic features of the overlapper were extracted. These results are shown in Table 2. We observe that, in order to achieve a performance of 32.9% EER (26.0% on the DEV set), our classifier would need 0.6s after the start of simultaneous speech to determine whether an overlap is COMP or COOP.

ACOUST_B_IN		
time after start overlap (in seconds)	DEV	TEST
1	27.5	33.5
0.8	26.7	34.4
0.6	26.0	32.9
0.4	32.4	36.4
0.2	31.1	40.0

Table 2: *Results of the* DEV *and* TEST *sets in % EER using acoustic features from the overlapper during overlap (*ACOUST_B_IN*).*

When we use the difference in acoustic behavior between the overlapper and overlappee as features, we find that the performances decrease, see Table 3 and 4. Apparently, this type of relative information is not sufficiently informative when used on its own. When these difference features *are* used, the preceding context of the overlappee seems to hold more discriminative information than the context during overlap.

DIFF_AB		
time after start overlap (in seconds)	DEV	TEST
1	34.7	43.6
0.8	36.2	42.2
0.6	34.7	38.6
0.4	36.7	41.4
0.2	39.6	45.1

Table 3: *Results of the* DEV *and* TEST *sets in % EER using the difference between the acoustic behavior of the overlapper (measured in overlap) and overlappee (measured in 1s-preceding context).*

Instead of using the DIFF_AB and DIFF_AB_IN features on their own, we also used them in addition to the ACOUST_B_IN feature set, as well as other features such as the head movement and focus of attention features. Table 5 and 6 show the results obtained when difference features, head movement and focus of attention features are added (by concatenating features on feature-level). It appears that the addition of FOA and HEAD features in some cases improve performance slightly. But it is not conclusive whether the preceding context or the in-overlap context is most beneficial.

DIFF_AB_IN		
time during overlap (in seconds)	DEV	TEST
1	45.7	54.1
0.8	41.2	49.3
0.6	39.4	47.1
0.4	39.2	45.1
0.2	39.1	41.4

Table 4: *Results of the* DEV *and* TEST *sets in % EER using the difference between the acoustic behavior of the overlapper and overlappee, both measured during overlap.*

Features	Dev	Test
ACOUST_B_IN	26.0	32.9
plus DIFF_AB	28.2	34.4
plus FOA_AB	27.7	30.1
plus HEAD_AB	25.6	32.7

Table 5: *Results when other features from the 1s-preceding context are added to the base* ACOUST_B_IN *feature set, measured during 0.6s overlap.*

Features	Dev	Test
ACOUST_B_IN	26.0	32.9
plus DIFF_AB_IN	26.7	30.7
plus FOA_AB_IN	24.2	27.9
plus HEAD_AB_IN	26.3	32.1

Table 6: *Results when other features are added to the base* ACOUST_B_IN *feature set, all measured during 0.6s overlap.*

5.2. Feature analysis

We take a closer look at how the extracted acoustic features differ from each other with respect to the distinction between COOP and COMP. Since the classification experiments show that this distinction is easiest made when features are extracted over a duration of 0.6s in overlap, we draw Box Whisker plots for all mean acoustic features measured over 0.6s in overlap, see Fig. 2. For most of the features we can observe differences between the COMPs' and COOPs' distributions of feature values. The largest differences can be found for INTENS, EFFORT, BREATHY1, and BREATHY2. In comparison to cooperative overlaps, competitive overlaps show higher values for intensity, vocal effort and the BREATHY2 measure, while the reverse direction is true for coarseness and the BREATHY1 measure. For the two features exhibiting the largest differences between COMP and COOP, i.e., the INTENS and EFFORT features, we visualize these features' behaviors of both the overlapper and the overlappee in the vicinity of speech overlap. We look at a 3s-long context between 2s prior to the start of an overlap and 1s after the start of an overlap, and align each overlap instance in our data by their starts. Subsequently, we aggregate the INTENS and EFFORT features of all overlap instances by averaging these values in the specified context. Although lumping the samples onto a big pile of data and subsequently taking the average may seem a coarse procedure, the common trend prevails which is what we would like to visualize with these plots. The contours of INTENS and EFFORT in the vicinity of overlap are shown in Fig. 3 and 4 respectively. As the visualizations and classification results show, it is mostly the absolute feature value of the overlapper during overlap that separates cooperative and competitive overlaps from each other, rather than the difference between the overlapper and overlappee.

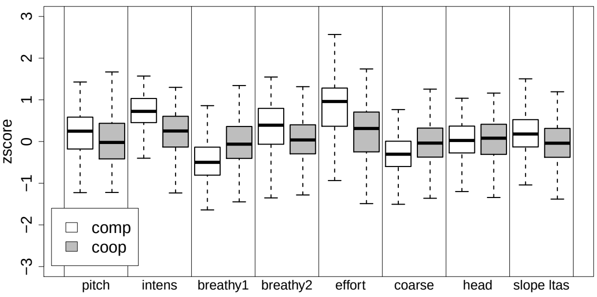
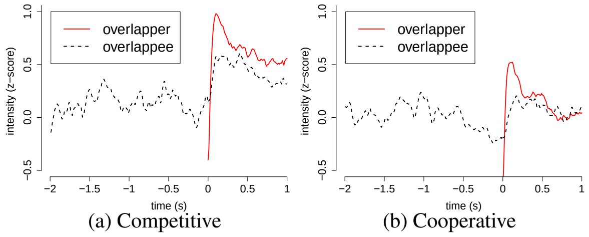
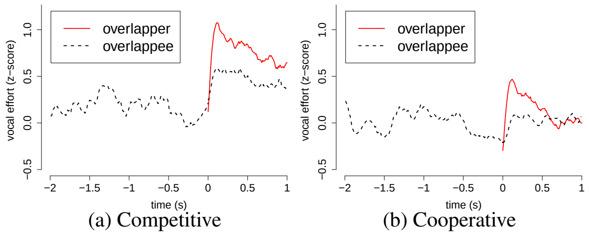


Figure 2: *Box Whisker plots of features (mean), measured over 0.6s in overlap.*



(a) Competitive (b) Cooperative

Figure 3: *Intensity (averaged over all overlap instances) in the vicinity of overlap. The start of overlap is at 0s.*



(a) Competitive (b) Cooperative

Figure 4: *Vocal effort (averaged over all overlap instances) in the vicinity of overlap. The start of overlap is at 0s.*

6. Conclusions and discussion

We developed classifiers for the classification of competitive and cooperative overlaps. We find that with a delay of 0.6s after the start of overlap, overlaps can be classified as competitive or cooperative with an EER of 32.0% (26.0% on dev set) using acoustic features of the overlapper only. Adding acoustic information from the overlappee by combining these features on feature-level does not improve performance. Slight improvement was obtained when gaze information during overlap was added. Visualizations of the overlappers' and overlappees' intensity and vocal effort contours in the proximity of overlap show significant differences between competitive and cooperative overlaps. Competitive overlaps show higher intensity levels, and higher levels of max energy in the mid-range frequency band of 0–2kHz than cooperative overlaps.

Given the performances obtained, there is much room for improvement. These improvements may lie in the use of other types of acoustic features or modelling techniques for which more annotated training data is needed. Finally, we suggest to further investigate how gazing behavior of both the overlapper and overlappee contribute to overlap classification.

7. Acknowledgements

This work was funded by the EU's 7th Framework Programme (FP7/2007-2013) under grant agreement no. 231287 (SSPNet).

8. References

[1] L. Yang, "Visualizing spoken discourse: prosodic form and discourse functions of interruptions," in *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, 2001, pp. 1–10.

[2] H. Z. Li, "Cooperative and intrusive interruptions in inter- and intracultural dyadic discourse," *Journal of Language and Social Psychology*, vol. 20, pp. 259–284, 2001.

[3] K. Murata, "Intrusive or cooperative? a cross-cultural study of interruption," *Journal of Pragmatics*, vol. 21, pp. 385–400, 1994.

[4] J. A. Goldberg, "Interrupting the discourse on interruptions," *Journal of Pragmatics*, vol. 14, pp. 883–903, 1990.

[5] E. Shriberg, A. Stolcke, and D. Baron, "Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, disfluencies, and overlapping speech," in *Proceedings of ISCA Tutorial and Research Workshop on Prosody in Speech Recognition and Understanding Prosody*, 2001, pp. 139–146.

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

[7] 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.

[8] C. Oertel, M. Wlodarczak, A. Tarasov, N. Campbell, and P. Wagner, "Context cues for classification of competitive and collaborative overlaps," in *Proceedings of Speech Prosody*, 2012, pp. 721–724.

[9] C.-C. Lee, S. Lee, and S. S. Narayanan, "An analysis of multimodal cues of interruption in dyadic spoken interactions," in *Proceedings of Interspeech*, 2008, pp. 1678–1681.

[10] C.-C. Lee and Narayanan, "Predicting interruptions in dyadic spoken interactions," in *Proceedings of ICASSP*, 2010, pp. 5250–5253.

[11] A. Gravano and J. Hirschberg, "A corpus-based study of interruptions in spoken dialogue," in *Proceedings of Interspeech*, 2012.

[12] E. Kurtic, G. J. Brown, and B. Wells, "Resources for turn competition in overlap in multi-party conversations: speech rate, pausing and duration," in *Proceedings of Interspeech*, 2010, pp. 2550–2553.

[13] J. Carletta, "Unleashing the killer corpus: experiences in creating the multi-everything AMI meeting corpus," *Language Resources and Evaluation*, vol. 41, pp. 181–190, 2007.

[14] P. Boersma and D. Weenink, "Praat, a system for doing phonetics by computer," *Glot International*, vol. 5, no. 9/10, pp. 341–345, 2001.

[15] B. Hammarberg, B. Fritzell, J. Gauffin, J. Sundberg, and L. Wedin, "Perceptual and acoustic correlates of abnormal voice qualities," *Acta Oto-laryngologica*, vol. 90, pp. 441–451, 1980.

[16] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.