INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Emotional Features for Speech Overlaps Classification

Olga Egorow, Andreas Wendemuth

Cognitive Systems Group, Otto von Guericke University, 39016 Magdeburg, Germany
olga.egorow@ovgu.de

Abstract

One interesting phenomenon of natural conversation is overlapping speech. Besides causing difficulties in automatic speech processing, such overlaps carry information on the state of the overlapper: competitive overlaps (i.e. "interruptions") can signal disagreement or the feeling of being overlooked, and cooperative overlaps (i.e. supportive interjections) can signal agreement and interest. These hints can be used to improve human-machine interaction. In this paper we present an approach for automatic classification of competitive and cooperative overlaps using the emotional content of the speakers' utterances before and after the overlap. For these experiments, we use real-world data from human-human interactions in call centres. We also compare our approach to standard acoustic classification on the same data and come to the conclusion, that emotional features are clearly superior to acoustic features for this task, resulting in an unweighted average f-measure of 71.9%. But we also find that acoustic features should not be entirely neglected: using a late fusion procedure, we can further improve the unweighted average f-measure by 2.6%.

Index Terms: speech overlaps, emotion recognition, prosodic features

1. Introduction

Speech overlaps – i.e. parts of a conversation where two or more participants speak simultaneously – occur more often than one would think. There are different findings on the frequency of overlaps in speech, depending on the kind of the interaction. In general, it is believed that approximately 40% of all between-speaker intervals are overlaps [1]. In telephone dialogues, the rate goes up to 52% [2]. The importance of overlaps lies in the fact that they can carry additional information. On one hand, overlaps can be seen as interruptions and are related to competition towards the other speaker [3], but on the other hand, they can also support the main speaker and be seen as signals of understanding [4]. Therefore overlaps can be distinguished into two types: cooperative and competitive. Since there are several different definitions available in the literature, we would like to provide a definition of overlaps that combines the most important aspects.

Cooperative overlaps are defined by the fact that the overlapper rather wants to support the current speaker than to interrupt him or her [5]. They are used to express supportive agreement or to complete an anticipated point [6]. In case of a cooperative overlap, the overlappee should not be offended [7]. The overlapper wants to maintain the conversation and has no intention to grab the floor by taking the turn [8]. There is also no disruption of the conversational flow and the intention of the overlapper is to keep attention on the main speaker's point.

Competitive overlaps are defined by the fact that the overlapper competes for speech time or topic, and wants to attract attention away from the current speaker [9] or to express disagreement [6] [8]. A competitive overlap is an attempt to steal the floor, and breaks the flow of the conversation [5]. The overlappee could perceive this overlap as problematic and offending [7].

Being cooperative or competitive, overlaps can reveal useful information on the nature of the interaction. The most interesting aspect about it is the occurrence of overlaps in spoken human-machine interaction, since overlaps are not an exclusive human-human interaction phenomenon, but occur in human-machine interaction, too [10]. To obtain this additional information, we first need to detect overlaps. There are different apporaches available in the literature, for instance using the overlappee's acoustic cues and the overlapper's gestural cues to predict the occurrence of overlaps, achieving an f-measure of 54% [11]. But more important than simply detecting the overlaps is to be able to automatically distinguish between cooperative and competitive overlaps. Here, we can also find some approaches in the literature. Most of them use only acoustic features, above all F_0, energy, intensity and loudness, since competitive overlaps are related to raising pitch and loudness [12] [13] [14] [15]. Some of the recent findings shall be presented below. It seems that the most promising approach is to use intensity-related features in combination with behavioural features like body movements and gaze. Oertel et al. use acoustic features (intensity and fundamental frequency-related) and body movement features, achieving a median accuracy of 63% [16]. Lee et al. combine intensity with body movements – in this case hand motions – obtaining a classification accuracy of 71.2% [17]. Another approach presented by Truong et al. uses the acoustic features of the overlapper only and achieves an EER of 32% for the cooperative vs. competitive classification task – which is slightly improved by adding multimodality through gaze information [18]. Among prosodic features, fundamental frequency and intensity were found to perform best using decision trees on overlap placement and acoustic features [19]. Chowdhury et al. investigate a variety of features - lexical, psycholinguistic and acoustic features – and achieve 66.43% classification accuracy using these features [4].

In this paper, we follow a different approach: we will show that the emotional content of both, the overlappee's and the overlapper's turns before and after the overlap can help distinguish between cooperative and competitive overlaps. We also will show that it is possible to classify competitive and cooperative overlaps using changes in valence and control levels of the utterances as features, and that this classification based on emotional labels outperforms the standard acoustic approach. Finally, we will compare the results obtained using only these emotional features to results using acoustic features, and results obtained employing early and late fusion procedures.

The paper is organised as follows: in Section 2 we describe the used data, Section 3 focuses on the two kinds of features we use, in Section 4 we introduce our experimental setup, Section 5 presents the achieved results, in Section 6 we discuss our findings, before we summarise our work in Section 7.

http://dx.doi.org/10.21437/Interspeech.2017-87

2. Data – The Davero Corpus

The Davero Corpus is a collection of real telephone-based human-human dialogues recorded in a German call centre. The calls are of different nature, from informational calls to complaints, including negative as well as positive emotions. The subset used in our experiments contains 47 dialogues recorded over 7 days. Besides speaker turns, the emotions uttered in the recordings are also annotated, based on the Geneva Emotion Wheel [20]. Turns with increasing control or valence are labelled by C+ and V+, respectively, turns with decreasing control or valence are labelled by C- and V-, respectively. More details on the Davero corpus and its annotation can be found in [21].

For the purpose of this investigation, we chose 47 random dialogues, as mentioned above. These dialogues contain 254 overlaps. The overlaps were labelled independently by two annotators – to ensure the high quality of the labels, we limited the data to only those overlaps, where the annotators agreed on the labels, resulting in 213 overlaps, 64 of which are competitive, and 149 cooperative. The overlaps are of both kinds: the call centre agent interrupting the client as well as the client interrupting the agent.

3. Used Features

For our experiments, we used different sets of features. These feature sets can be divided into two kinds: acoustic and emotional.

3.1. Acoustic features

The first acoustic feature set is the *emobase* set, well-known in the community of emotion recognition from speech. It contains 988 acoustic features consisting of functionals (such as mean, standard deviation, min, max, etc.) of low-level acoustic descriptors (such as MFCCs, loudness, voicing probability, intensity, etc.) and their derivatives. The features are extracted on utterance level, resulting in one data point per utterance. The full description of this feature set can be found in [22]. We chose this feature set because of its good performance in a variety of problems in the field of acoustic emotion recognition. We will refer to this feature set as A_1.

The second acoustic feature set is constructed according to the state of the art: since several sources use features based on intensity and fundamental frequency, we built a subset of the emobase feature set containing functionals only of these two features, resulting in 38 features (19 intensity-related and 19 fundamental frequency-related). We will refer to this feature set as A_2.

Since the audio streams of the agent and the client are not separated in our data, we decided to extract acoustic features of the utterance in which the overlap occurs.

3.2. Emotional features

The emotional features are based upon the emotional labels available for the Davero Corpus. All the utterances of the agent and the client are labelled according to the emotions expressed by the speaker in a two-dimensional emotional space, the two dimensions being control and valence. The utterances with changes of the control or valence level are annotated with C+, V+ for increasing control and valence, and C- and V- for decreasing control and valence, respectively. The utterances without labels do not contain control or valence level changes.

For this investigation, we considered the four turns surrounding the overlap, two turns before and two turns after, resulting in three feature sets. Feature set E_1 contains features from the two turns before the overlap (*turn-2* and *turn-1*), resulting in 4 features: control in turn-2, valence in turn-2, control in turn-1, valence in turn-1. Feature set E_2 contains features from the two turns after the overlap (*turn+1* and *turn+2*): control in turn+1, valence in turn+1, control in turn+2, valence on turn+2. Feature set E_3, finally, contains all eight features. An example of an annotated overlap is shown in Fig. 1. The extracted features for the depicted case are, according to the scheme described above, [+1, 0, -1, -1, +1, -1, +1, +1, +1].

4. Experimental Setup

The first step in the processing pipeline was data partitioning. In our setting, a speaker-independent evaluation posed a challenge: The data contained recordings of the same five agents interacting with 47 clients. Since the conversations differ significantly regarding the length and the number of contained overlaps, we decided not to do a leaving-one-speaker-out setting but to divide the data into days, resulting in a training data set containing recordings of five days and a test set containing recordings of the remaining two days. The recordings of different days still differ in length and number of contained overlaps and their kind, but we tried to keep the training and test data sets as similar as possible. The distribution of overlaps and speakers can be found in Table 1.

Table 1: *Distribution of overlaps and speakers*

Overlaps	Coop	Comp
Training data	116	53
Test data	33	11
Speakers' sex	Agents	Clients
Training data	3 male : 2 female	12 male : 24 female
Test data	1 male : 1 female	7 male : 6 female

We conducted the classification experiments using the Support Vector Machine (SVM) implementation provided by the libSVM library [23] in KNIME [24]. This implementation also provided the probability estimates that we used for our late fusion procedure described in Section 5. To fine-tune the parameters of the classifier, we used a development data set, consisting of a randomly chosen 10%-subset of the training data set. Using this setup, we tested different kernels (linear, polynomial and radial basis function) and soft margin parameters C. The best results were achieved using a linear kernel and C = 10.

5. Results

As already mentioned in Section 3, we use acoustic features and emotional features. Therefore, we conducted the classification experiments first on acoustic and emotional features separately, and then tested early as well as late fusion procedures.

5.1. Acoustic features

The results achieved using the two acoustic feature sets mentioned earlier are shown in Table 2, first using the complete emobase set (A_1) and then using only intensity and fundamental frequency related features (A_2). We can see that the results using these feature sets are only 2.5% to 8.1% above chance level (52.5% and 58.1% unweighted average f-measure, respec-

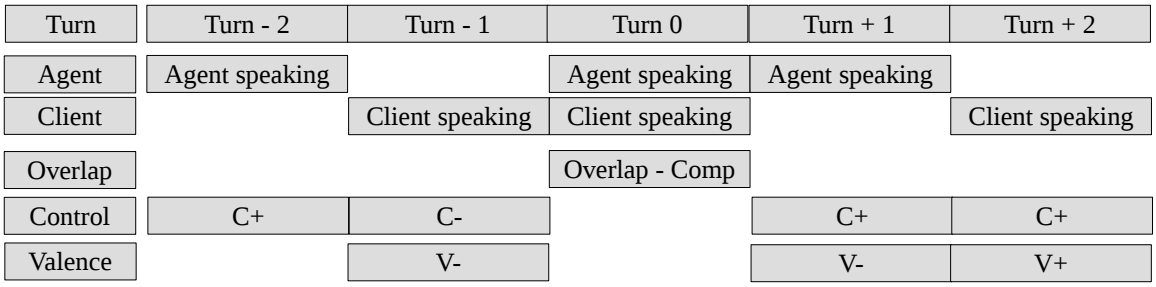

Figure 1: *The annotation of an exemplary overlap and its surroundings. The annotation consists of five tiers: the agent and the client tiers containing their turns, the overlap tier containing the label of the overlap, and two emotional annotation tiers: the control and the valence tier.*

tively). A_2 performs slightly better, with both the unweighted average recall and precision being higher than using A_1. These low results are induced by the low performance for the competitive class – instances of the cooperative class can be found with a relatively high recall and precision of 69.7% / 76.7% and 87.9% / 78.4% for A_1 and A_2, respectively. This imbalance is most likely caused by the unbalanced class distribution, with the cooperative class appearing more than twice as frequently in the data than the competitive class. Another cause for the suboptimal performance of the acoustic features can be found in the unbalanced sex distribution of the data (twice as many female speakers as male speakers on the training data), since the speakers' sex is known to have an influence on the recognition of affective states from speech [25].

Table 2: *Classification results of acoustic features on the test set in terms of recall, precision and f-measure.*

Emobase (A_1)	Recall	Precision	F-Measure
Coop	0.697	0.767	0.730
Comp	0.364	0.286	0.320
UA	0.530	0.526	0.525
Int + F0 (A_2)	Recall	Precision	F-Measure
Coop	0.879	0.784	0.829
Comp	0.273	0.429	0.333
UA	0.576	0.606	0.581

5.2. Emotional features

As already mentioned in Section 3, in the case of emotional features, we considered two turns before and after the overlap to classify the nature of the overlap. Table 3 presents the achieved results when using only the turns before the overlap (E_1), only the turns after the overlap (E_2), and both, turns before and after the overlap (E_3).

In contrast to the results achieved using acoustic features, the results obtained here are clearly above chance level. The best performance is achieved by using E_3, resulting in an unweighted average f-measure of 71.9%, followed by the performance of E_2 with 69.3% unweighted average f-measure.

5.3. Fusion

Since different modalities can contain additional information, using multimodal data can improve the results of automatic classification – almost all approaches mentioned in Section 1 use different kinds of features. A good overview over multimodal fusion techniques can be found in [26].

Table 3: *Classification results on emotional features – uttered in turns only before the overlap (E_1), only after the overlap (E_2), and in both, before and after the overlap (E_3)*

E_1	Recall	Precision	F-Measure
Coop	0.848	0.824	0.836
Comp	0.455	0.500	0.476
UA	0.652	0.662	0.656
E_2	Recall	Precision	F-Measure
Coop	0.788	0.867	0.825
Comp	0.636	0.500	0.560
UA	0.712	0.683	0.693
E_3	Recall	Precision	F-Measure
Coop	0.879	0.853	0.866
Comp	0.545	0.600	0.571
UA	0.712	0.726	0.719

To prove this hypothesis for our case, we tested two different fusion procedures: early fusion and late fusion. In the first case, we fused the acoustic and emotional features and trained the classifier on them simultaneously. In the second case, we fused the classification results obtained using only acoustic and only emotional features. For this, we used the probability estimates provided by libSVM of both acoustic and emotional classifiers, and calculated the mean value. We obtained the class labels from this calculated probability mean value by setting the threshold at 0.5.

The results of both fusion procedures are shown in Table 4. For both procedures, we decided to fuse the feature set A_2 and the feature set E_3, since they provided the best results when used alone. We can see that, by achieving 69.7% unweighted average f-measure, early fusion does not reach the results obtained by using only emotional features, although it clearly outperforms using only acoustic features. Late fusion, on the other hand, does provide an improvement over using only one modality, outperforming acoustic features as well as emotional features. Although the improvement of 2.6% absolute unweighted average f-measure may not seem significant, this procedure improves the precision of the classification by 11.2% absolute compared to emotional features (the precision increases from 72.6% to 83.8%).

Table 4: *Classification results on fused features (A_2 + E_3)*

Early fusion	Recall	Precision	F-Measure
Coop	0.848	0.848	0.848
Comp	0.545	0.545	0.545
UA	0.697	0.697	0.697
Late fusion	Recall	Precision	F-Measure
Coop	0.970	0.842	0.901
Comp	0.455	0.833	0.588
UA	0.712	0.838	0.745

6. Discussion

The most interesting result of this investigation is, without doubt, the fact that the competitive or cooperative nature of an overlap can be distinguished using the emotional content of the turns before and after the overlap. From this finding we can conclude that the emotional content of the utterances surrounding an overlap does differ depending on the nature of the overlap – especially the emotions after the overlap, since the feature set containing these features performed better than the one containing features from before the overlap. From the achieved results we can also see that using emotional content of the utterances as a modality outperforms most approaches described in the literature.

Another interesting finding is that, although using only acoustic features does not provide useful results, adding them as an extra modality improves the results obtained on emotional features by 2.6% absolute in terms of f-measure and 11.2% absolute in terms of precision. This can be interpreted as a hint that acoustic features do provide additional information compared to only the emotional features.

The usage of emotional features, however, poses a problem – the extraction of emotional features is not as effortless as the extraction of acoustic features, that can be easily done using standard software like openSMILE. For our investigations, we used manually labelled emotions – but this procedure is not applicable in real-world scenarios. This problem can be solved by using automatic emotion recognition methods to detect changes of the control and valence levels – and there are approaches that deliver acceptable results, e.g. an approach for the Davero data set used in our experiments [27]. Using automatically extracted emotional labels might impair the overlap classification results. This, as well as introducing automatic utterance segmentation, needs to be addressed in further investigations in order to make the approach applicable to real-world scenarios without further manual interference.

7. Conclusion

In this paper, we have shown that competitive and cooperative overlaps in spontaneous, telephone-based human-human conversations can be classified using the emotional content of the utterances surrounding the overlap. We also compared the classification performance using emotional features to performance using acoustic features and found that emotional features are clearly superior. We achieved an unweighted average f-measure of 71.9% using the emotional content of the four utterances before and after the overlap. We further improved this result by fusing it with results obtained on acoustic features, resulting in 74.5% unweighted average f-measure. Since we use different data bases, we can hardly compare our result to other approaches in the literature, but in terms of numbers, our approach outperforms several of them, including those using multimodal features. Nevertheless, it is an interesting question for future research to compare the state-of-the-art approaches on a benchmark data set.

8. Acknowledgements

The authors thank for continued support by the SFB/TRR 62 “Companion-Technology for Cognitive Technical Systems” (www.sfb-trr-62.de) funded by the German Research Foundation (DFG). Further, this work was sponsored by the German Federal Ministry of Education and Research (BMBF) in the program Zwanzig20 Partnership for Innovation as part of the research alliance 3Dsensation (www.3d-sensation.de) under grant number 03ZZ0414.

9. References

[1] M. Heldner and J. Edlund, “Pauses, gaps and overlaps in conversations,” *Journal of Phonetics*, vol. 38, no. 4, pp. 555–568, 2010.

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

[3] C. West, “Against our will: Male interruptions of females in cross-sex conversation,” *Annals of the New York Academy of Sciences*, vol. 327, no. 1, pp. 81–96, 1979.

[4] A. Chowdhury, M. Danieli, and G. Riccardi, “The role of speakers and context in classifying competition in overlapping speech,” in *Proc. of INTERSPEECH-2015*, 2015, pp. 1844–1848.

[5] K. Murata, “Intrusive or co-operative? a cross-cultural study of interruption,” *Journal of Pragmatics*, vol. 21, no. 4, pp. 385–400, 1994.

[6] L.-c. Yang, “Visualizing spoken discourse: Prosodic form and discourse functions of interruptions,” in *Proc. of the Second SIGdial Workshop on Discourse and Dialogue-Volume 16*. Association for Computational Linguistics, 2001, pp. 1–10.

[7] S. A. Chowdhury, M. Danieli, and G. Riccardi, “Annotating and categorizing competition in overlap speech,” in *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2015, pp. 5316–5320.

[8] H. Z. Li, “Cooperative and intrusive interruptions in inter-and intracultural dyadic discourse,” *Journal of Language and Social Psychology*, vol. 20, no. 3, pp. 259–284, 2001.

[9] J. A. Goldberg, “Interrupting the discourse on interruptions: An analysis in terms of relationally neutral, power-and rapport-oriented acts,” *Journal of Pragmatics*, vol. 14, no. 6, pp. 883–903, 1990.

[10] I. Siegert, R. Bock, A. Wendemuth, B. Vlasenko, and K. Ohnemus, “Overlapping speech, utterance duration and affective content in hhi and hci – an comparison,” in *Cognitive Infocommunications (CogInfoCom), 2015 6th IEEE International Conference on*. IEEE, 2015, pp. 83–88.

[11] C.-C. Lee and S. Narayanan, “Predicting interruptions in dyadic spoken interactions.” in *ICASSP*, vol. 10, 2010, pp. 5250–5253.

[12] P. French and J. Local, “Turn-competitive incomings,” *Journal of Pragmatics*, vol. 7, no. 1, pp. 17–38, 1983.

[13] 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.

[14] B. Hammarberg, B. Fritzell, J. Gaufin, J. Sundberg, and L. Wedin, “Perceptual and acoustic correlates of abnormal voice qualities,” *Acta oto-laryngologica*, vol. 90, no. 1-6, pp. 441–451, 1980.

[15] E. Shriberg, A. Stolcke, and D. Baron, “Can prosody aid the automatic processing of multi-party meetings? evidence from predicting punctuation, dis uencies, and overlapping speech,” in *ISCA Tutorial and Research Workshop (ITRW) on Prosody in Speech Recognition and Understanding*, 2001.

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

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

[18] K. P. Truong, "Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee," in *Proc. of INTERSPEECH-2013*. International Speech Communication Association, 2013, pp. 1404–1408.

[19] E. Kurtić, G. J. Brown, and B. Wells, "Resources for turn competition in overlapping talk," *Speech Communication*, vol. 55, no. 5, pp. 721–743, 2013.

[20] K. R. Scherer, "What are emotions? And how can they be measured?" *Social science information*, vol. 44, no. 4, pp. 695–729, 2005.

[21] I. Siegert, D. Philippou-Hübner, M. Tornow, R. Heinemann, A. Wendemuth, K. Ohnemus, S. Fischer, and G. Schreiber, "Ein Datenset zur Untersuchung emotionaler Sprache in Kundenbindungsdialogen," in *Proc. of the 26th ESSV, Eichstätt, Germany*, 2015, pp. 180–187.

[22] F. Eyben, M. Wöllmer, and B. Schuller, "openSMILE - The Munich Versatile and Fast Open-Source Audio Feature Extractor," in *Proc. of the ACM MM-2010*, Firenze, Italy, 2010, pp. 1459–1462.

[23] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 2, pp. 1–27, 2011.

[24] M. R. Berthold, N. Cebron, F. Dill, T. R. Gabriel, T. Kötter, T. Meinl, P. Ohl, C. Sieb, K. Thiel, and B. Wiswedel, "KNIME: The Konstanz Information Miner," in *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL 2007)*. Springer, 2007.

[25] I. Siegert, D. Philippou-Hübner, K. Hartmann, R. Böck, and A. Wendemuth, "Investigation of speaker group-dependent modelling for recognition of affective states from speech," *Cognitive Computation*, vol. 6, no. 4, pp. 892–913, 2014.

[26] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," *Multimedia systems*, vol. 16, no. 6, pp. 345–379, 2010.

[27] I. Siegert and K. Ohnemus, "A new dataset of telephone-based human-human call-center interaction with emotional evaluation," in *Proc. of the First International Symposium on Companion Technology (ISCT)*, 2015.