CLASSIFYING OVERLAPPEE AND OVERLAPPER BASED ON ACOUSTIC PROSODIC FEATURES

Maximilian Nothnagel[1], Barbara Schuppler[1]

[1] *Signal Processing and Speech Communication Laboratory, Graz University of Technology, Austria*

This paper is a revised version of the authors' submitted manuscript that was peer-reviewed by at least two anonymous reviewers.

Abstract

Conversational overlaps, i.e., instances where two or more participants speak simultaneously, are a prevalent feature of human communication, influencing conversational dynamics and speaker roles. While previous studies have mainly analysed and categorised them based on competitiveness, this paper focuses on the roles of individual speakers within these overlaps. Our study is based on instances of overlapping speech from a corpus of casual, spontaneous dyadic conversations, from which we extract acoustic features and communicative function labels that were created manually. Using XGBoost for classification, we find that overlaps as coming from an overlappee or overlapper, can be discriminated with a balanced accuracy score of approx. 75 %. Our analysis of feature importances shows that the communicative function along with the intensity levels are the key indicators in identifying speaker roles during overlaps. This study contributes to our understanding of conversational dynamics and provides insights applicable to making turn-taking in human-computer interaction more natural.

Keywords: *overlapping speech, conversational speech, turn-taking, acoustic prosodic features*

1. Introduction

Spontaneous speech exhibits the natural phenomenon of overlapping speech. That is, more than one interlocutor speaking at the same time for any duration during a conversation. This can happen through backchannelling to convey agreement or interruptions to compete for the role of the foreground speaker [1], [2]. It has been shown that overlapping speech is not affected by the familiarity of the speakers, nor whether the conversation is of a multi-party or dyadic nature [1], [3]. In addition, the existence—and usage—of overlap has an impact on various areas of speech processing, for example on the error rate of speaker diarisation in meeting room conversations [2], [4].

Overall, it has been shown that overlaps occur in various corpora, dyadic and multi-party, albeit to a different degree. For instance, the dyadic task-oriented Columbian Games Corpus [5], with a total of nine hours of orthographically transcribed dialogue, contains 1 329 segments of overlap [6]. In [7], it was shown that from 27 hours of call centre conversation, 9 858 overlap segments were obtained for a duration of approximately four hours. In the ISCI Meeting Project [8], it was shown that 8,8–17 % of words and 31,4–54,5 % of spurts were overlapped [1], [4]. Based on a total of 9 816 segments of a subset of this corpus, [9] showed that 46 % were overlap instances. In [10], five meetings from the multi-party AMI Meeting corpus [11] were shown to contain in total 509 overlap segments.

Related work focused on few major areas. The first being classification of the competitiveness of overlaps with respect to turn-taking. The most common approach is differentiating competitive versus non-competitive using acoustic features, such as [7] with low-level features like pitch, loudness, energy, or voice-quality, among others. [9] utilised (besides prosodic features of speakers) contextual data, pausing, duration and placement of the overlap, and turn completion. In [10], head movement and speaker's gaze was additionally incorporated to identify competitive and cooperative overlaps. More granular categories of competitiveness were tested in [12], but rejected for the binary classification using (contextual) frequency features and overlap onset types.

Other research areas include the detection and utilisation of overlap data in speaker diarisation [2], [4] or the characterisation of speech during interruptions [6]. It is common notion to define the participants in an overlap as overlappee (EE) and overlapper (ER). This definition is based on whose current turn unit started prior [12], but can be interpreted as who initiated (overlapper) the overlap and who did not (overlappee) [13].

Corresponding author: *b.schuppler@tugraz.at.*

In this paper, we investigate whether the acoustic information in the overlap and surrounding context for any overlap participant contains sufficient information—despite having a rather small dataset, such that it can be automatically classified whether those instances of speech comes from the overlappee or overlapper. By using prosodic and durational features along with labels representing the communicative function in turn-taking, we do not only build a classification model but also gain insights with respect to the relative importance of our features to the disambiguation of overlappee and overlapper. In a second classification experiment, we explore the feasibility of deducing the communicative function of the speaker's turn from the acoustic characteristics of the overlapping speech.

2. Materials and Methods

2.1. GRASS corpus

For the analysis of overlaps, we use the Graz corpus for Read and Spontaneous Speech (GRASS) [14], [15], a speech corpus containing spontaneous dyadic conversations. For one hour, 19 pairs of adult speakers of Austrian German who knew each other well prior to the recordings talked about self-chosen every-day topics, while no experimenter was in the room. The speakers were recorded with close talking head-sets (AKG HC-577L) on separate channels in a high-quality recording studio.

All conversations were manually transcribed orthographically using Praat [16]. Additionally, 300 seconds of turn-taking annotations were annotated manually in the form of inter-pausal units (IPUs) (i.e., Inter Pausal Units with a 150ms criterion) and point of potential completion (PCOMP) annotations (i.e., points of potential completion in an utterance, indicating possible turn-taking opportunities based on syntactic criteria) on separate tiers for 14 conversations (i.e., 28 speakers, male and female) [17].

2.2. Overlap Extraction

The overlap detection from these 14 conversations was done with TextGridTools [18], using orthographic transcriptions and forced alignment based word boundaries [19]. Only overlapping talk was considered, so any overlap caused by *laughter*, *breathing*, or similar was ignored. Utilising the precise boundaries and timestamps from the phonetic segmentation, we determined the *overlap type* for each interlocutor of an overlap as either overlappee (EE) or overlapper (ER).

For each speaker in an overlap, we obtained the PCOMP label of the respective turn as a categorical *turn-taking* feature. If an overlap spanned over multiple labels, the overlap was split into multiple data entries based on the manually annotated turn-taking intervals. For further processing and simplification, the PCOMP labels were, as in [20], grouped to either `hold` (i.e., all labels indicating that the same speaker continues speaking), `change` (i.e., all labels indicating that the floor is given to the interlocutor), or `hrt` (i.e., a label for backchannels and other (also) lexical tokens providing feedback without yielding the floor), as shown in Table 1. Note that combined labels, e.g. `hold_question`, were prior grouped with their first label. This is an acceptable choice according to [17], but leads to an interchange of 47 and 30 labels from `hold` to `change`, and vice-versa, due to `change` and `question` being mostly combined with labels from the `hold` group.

Table 1: Grouping of PCOMP labels in the dataset.

Original label	Grouped label

`hold`
`cont`
`part`
`q-part`
`hes`
`coll`
`disruption`
	`hold`

`change`
`question`
`incomplete`
	`change`

`hrt`	`hrt`

Table 2: Number of grouped PCOMP labels per overlap type.

	Overlap Type		
PCOMP Label	EE	ER	Total
`hold`	299	236	535
`change`	216	126	342
`hrt`	65	215	280
NA	19	22	41

In total, 1 198 data entries were obtained from 598 overlaps, varying from 7 to 120 overlap pairs per conversation. The numbers of specific PCOMP labels per overlap type are shown in Table 2, already hinting at tendencies of EE for `change`, ER for `hrt`, and an (almost) equal distribution of `hold` labels. Overall, 41 turns were marked as uncertain or unable to be determined.

In addition, two *durational* features were extracted. Based on the word boundaries of both speakers, we computed the duration of the turn-based overlap. Using this the percentage of the speaker's turn in the overlap was calculated as a relative overlap time feature.

2.3. Acoustic prosodic feature extraction

Similar to [21], pitch tracking and frequency extraction were performed on the intervals using pYAAPT [22], a Python implementation of Yet Another Algorithm for Pitch Tracking (YAAPT) [23], and peak detection using detecta [24]. Frequency (F0) values were extracted with pYAAPT's default configuration, i.e. analysis frame length of 35 ms and frame spacing of 10 ms. Intensity (RMS) values were calculated using a block processing algorithm with a window time of

25 ms and a step time of 10 ms.
To enhance stability for shorter segments, we added a buffer of 0,4 s to both ends of the source time frame. This was exclusively applied during pitch tracking only, the values corresponding to the original timestamps were extracted for further processing. The values were normalised using the speaker's respective mean and standard deviation to transform them to z-scores [10], [12].
We extracted F0 and RMS values from the overlapping speech and its context, separately. As context, we considered either the immediate neighbouring turn, if it exists, or otherwise the previous/subsequent remaining part of the overlapping turn, EE and ER, respectively. If an overlapping turn has no immediate neighbours and is fully in overlap, there is no context. For each data entry, we computed mean, standard deviation, minimum valleys, and maximum peaks for the normalised F0 and RMS values. In addition, the steepness per sample of the linear polynomial fit on the value's curve was computed as slope for both features [12]. Due to the significant amount of missing values from peak detection, minimum and maximum features were dropped. This results in six *acoustic* features.
Furthermore, six *delta* features were calculated by comparing each acoustic feature of the overlap with the corresponding feature of the contextual time frame, i.e. $\delta_{\mu,\mathrm{RMS}} = \mu_{\mathrm{RMS}}^{\mathrm{ov}} - \mu_{\mathrm{RMS}}^{\mathrm{ctx}}$.

2.4. Classification

For classification, we used PyCaret [25] to predict overlap types and PCOMP labels with an Extreme Gradient Boosting (XGBoost) classifier. The model was chosen from a stratified 10-fold cross-validation using a 70 % train-test-split.
No imputation was performed, all rows with missing values were dropped. Constant imputation for delta features was tested, but not further pursued due to negligible changes in performance, resulting in the removal of circa 30 rows.
In total, this leaves us with a dataset of 1 126 data entries. The full data set contains 16 features, including the two respective targets, split into five categories, as explained in Table 3. For the binary classification of overlap types, PCOMP labels were one-hot encoded.

3. Results

3.1. Discriminating EE and ER

We use balanced accuracy (BACC) and F_1 scores as performance metrics. Furthermore, to gain insights with respect to the impact of the features, we look at the features' importances and SHapley Additive exPlanations (SHAP) values [26].
Besides the full feature set, we construct two additional sets: The first omits the PCOMP annotations to analyse the dependency of the model on the labels. The second is based on insights gained from the full feature set and contains only the six most important features (delta RMS values, RMS slope, and both temporal values), as well as the PCOMP label.

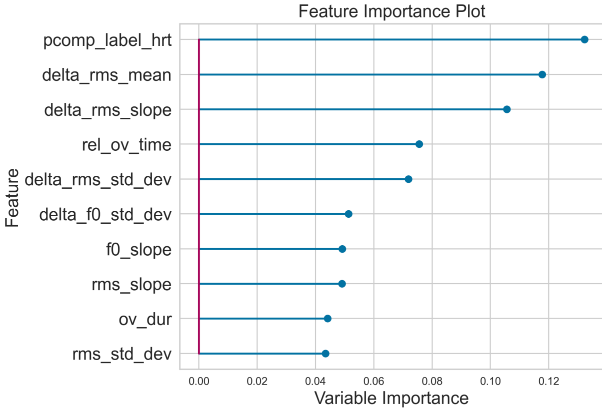

Figure 1: Feature importances for overlap type classification on the full feature set. Only ten most important features shown.

The performance metrics per feature set on the holdout data are shown in Table 4. The ten highest importance scores on the full feature set are shown in Fig. 1, the computed SHAP value per feature of each data entry in Fig. 2, highlighting which feature realisations drive the model to the overlap types. Furthermore, the mean absolute SHAP values per feature and overlap type are shown in Table 5 for the reduced feature set.

3.2. Classifying communicative functions

For the multi-class classification of the three PCOMP labels, the model performs comparatively poorly. On the holdout set, it achieves a BACC of 49,96 % and F_1 scores of 61,45 %, 28,22 %, and 58,82 % (`hold`, `change`, `hrt` respectively). Due to the mediocre performance, we did not further pursue the classification nor any improvements.

4. Discussion

Overall, the XGBoost classifier demonstrates reasonable efficacy in differentiating between the overlappee and the overlapper with a balanced accuracy score of circa 75 %. We show that F0 features and acoustic non-delta features are weighted as less important and show no strong impact on the classifier's output. In contrast, contextual RMS values, especially mean and slope, show some interpretable relation: An overlapper, mostly at the start of their overlapping turn, speaks with an "increased" loudness compared to the other person *and* themselves if they continue talking after the overlap. Speaking at roughly the same intensity, or even less, indicates still a slight tendency towards EE, but less noticeable.
In addition, if the overlap duration is relatively small, or a large portion of the turn is in overlap, the model tends strongly towards ER. Considering the high count of `hrt` in ER, this seems logical, as this type of turn tends, in general, to be shorter. On the other hand, an overlappee's turn tends to last longer until the (comparatively short) overlap occurs.

Table 3: Features available in the full feature set, including target(s).

Category	Features	# Features
Overlap Type	Overlappee (EE) or overlapper (ER)	1
Turn-Taking	PCOMP label of overlapping turn	1 †
Durations	Duration of overlap, and percentage of turn that is in this overlap	2
Acoustics	Computed mean, standard deviation, and slope for F0 and RMS	6
Deltas	Same features as acoustics, computed as e.g. $\mu^{\text{ov}}_{\text{RMS}} - \mu^{\text{ctx}}_{\text{RMS}}$	6

† As a feature, the label is one-hot encoded into 3 binary exclusive features.

Table 4: Performance metrics per feature set on the holdout data.

		F_1 Score	
Feature Set	BACC	EE	ER
Full	75,06 %	76,40 %	73,75 %
w/o PCOMP	76,41 %	77,22 %	75,60 %
Reduced	75,95 %	77,31 %	74,61 %

Table 5: Mean absolute SHAP values per feature and overlap type on the reduced feature set, sorted by sum of SHAPs.

	Mean Abs. SHAP	
Feature Name	EE	ER
`delta_rms_slope`	1,35	1,69
`rel_ov_time`	0,82	0,96
`delta_rms_mean`	0,87	0,89
`rms_slope`	0,70	0,98
`ov_dur`	0,79	0,80
`delta_rms_std_dev`	0,78	0,75
`pcomp_label_hrt`	0,19	0,32
`pcomp_label_change`	0,27	0,23
`pcomp_label_hold`	0,06	0,07

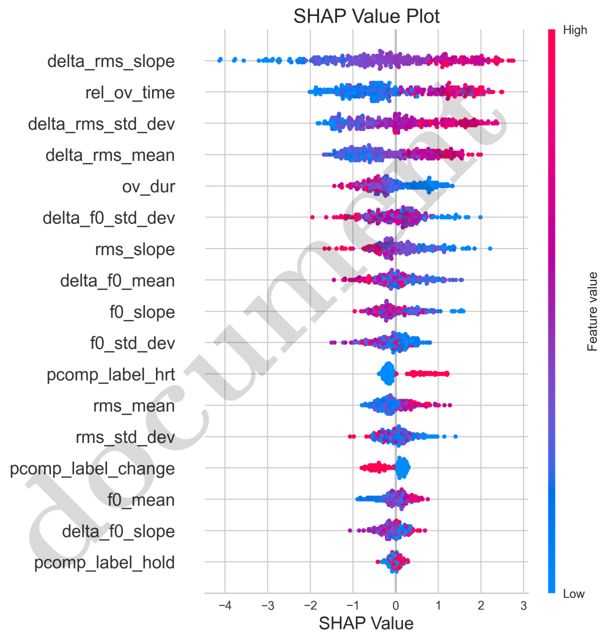

Figure 2: SHAP values per feature computed on the full feature set. Negative SHAP drives the output towards EE, positive to ER.

Regarding PCOMP annotations themselves, `change` and `hrt` have a noticeable, if slightly minor, impact, shown through the high importance for `hrt` and the clear distinction in SHAP values. Omitting this feature shows that the model does not strictly depend on the label, as its performance metrics remain largely unchanged, varying slightly by less than three percentage points. This might hint to uncertainties induced by `hold` or atypical `change` and `hrt` labels.
In addition, we construct a reduced feature set, containing the six most relevant features and the PCOMP label. Despite the reduced feature complexity, the reduced set shows slight improvement. In addition, we see that, again, `hold` has little impact on either target. The other two labels show clearer tendencies, but are still not as relevant as the acoustic and duration features, most prominently both RMS slope features and relative overlap time for ER.

5. Conclusion

In this paper, we demonstrated the potential of tree-based classifiers to differentiate between overlappee and overlapper using acoustic prosodic and temporal features, together with categorical turn-taking labels available within and surrounding the overlap. While certain turn-taking categories tended to have a more significant impact on the classification performance than others, the label itself showed less importance than the durational and acoustic prosodic features. Our classification experiments on determining the turn-taking label as target based on the acoustic features derived from the overlap performed poorly. Given our small data set, we can not draw conclusions yet about whether this low performance is caused by an insufficient amount of data points for each turn-taking category, or whether indeed the acoustic features of the overlap do not differ enough to disambiguate the PCOMP labels.
To conclude, our work contributes to overlap analysis, including the consideration of additional features, such as speech rate or overlap onset time, incremental feature selection approaches, and more granular context detection to prevent overly long context windows or influences from other overlaps. Our findings

may inform future studies on conversational dynamics and presents relevant insights applicable to human-computer interaction, where the timing in turn-taking still poses significant challenges.

6. Acknowledgements

This research was funded in part by the Austrian Science Fund (FWF) [10.55776/P32700].

References

[1] E. Shriberg, A. Stolcke, and D. Baron, "Observations on Overlaps: Findings and Implications for Automatic Processing of Multi-Party Conversation," in *Interspeech 2001*, Citeseer, 2001, pp. 1359–1362.

[2] S. H. Yella and H. Bourlard, "Overlapping Speech Detection Using Long-Term Conversational Features for Speaker Diarization in Meeting Room Conversations," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1688–1700, 2014. DOI: `10.1109/TASLP.2014.2346315`.

[3] E. Shriberg, "Spontaneous Speech: How People Really Talk and Why Engineers Should care," in *Interspeech 2005*, Citeseer, 2005, pp. 1781–1784.

[4] S. Otterson and M. Ostendorf, "Efficient use of overlap information in speaker diarization," in *2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, 2007, pp. 683–686. DOI: `10.1109/ASRU.2007.4430194`. [Online]. Available: `https://ieeexplore.ieee.org/document/4430194/` (visited on 05/10/2025).

[5] A. Gravano and J. Hirschberg, "Turn-taking cues in task-oriented dialogue," *Computer Speech & Language*, vol. 25, no. 3, pp. 601–634, 2011, ISSN: 0885-2308. DOI: `10.1016/j.csl.2010.10.003`. [Online]. Available: `https://www.sciencedirect.com/science/article/pii/S0885230810000690` (visited on 05/11/2025).

[6] A. Gravano and J. Hirschberg, "A Corpus-Based Study of Interruptions in Spoken Dialogue," in *Interspeech 2012*, 2012, pp. 855–858. DOI: `10.21437/Interspeech.2012-193`.

[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)*, 2015, pp. 5316–5320. DOI: `10.1109/ICASSP.2015.7178986`.

[8] N. Morgan, D. Baron, J. Edwards, *et al.*, "The Meeting Project at ICSI," in *Proc. of the 1st International Conference on Human Language Technology Research*, 2001. [Online]. Available: `https://aclanthology.org/H01-1051/` (visited on 05/19/2025).

[9] 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, ISSN: 0167-6393. DOI: `https://doi.org/10.1016/j.specom.2012.10.002`.

[10] K. P. Truong, "Classification of cooperative and competitive overlaps in speech using cues from the context, overlapper, and overlappee," in *Interspeech 2013*, 2013, pp. 1404–1408. DOI: `10.21437/Interspeech.2013-368`.

[11] J. Carletta, S. Ashby, S. Bourban, *et al.*, "The AMI Meeting Corpus: A Pre-announcement," en, in *Machine Learning for Multimodal Interaction*, vol. 3869, Springer Berlin Heidelberg, 2005, pp. 28–39, ISBN: 978-3-540-32550-5. DOI: `10.1007/11677482_3`. [Online]. Available: `http://link.springer.com/10.1007/11677482_3` (visited on 05/19/2025).

[12] E. Kurtić and J. Gorisch, "F0 accommodation and turn competition in overlapping talk," *Journal of Phonetics*, vol. 71, pp. 376–394, 2018.

[13] S. A. Chowdhury, M. Danieli, and G. Riccardi, "The Role of Speakers and Context in Classifying Competition in Overlapping Speech," in *Interspeech 2015*, ISCA, 2015, pp. 1844–1848. DOI: `10.21437/Interspeech.2015-68`. [Online]. Available: `https://www.isca-archive.org/interspeech_2015/chowdhury15_interspeech.html` (visited on 05/10/2025).

[14] B. Schuppler, M. Hagmüller, J. A. Morales-Cordovilla, and H. Pessentheiner, "GRASS: The Graz corpus of Read And Spontaneous Speech," in *Proc. of LREC*, 2014, pp. 1465–1470.

[15] B. Schuppler, M. Hagmüller, and A. Zahrer, "A Corpus of Read and Conversational Austrian German," *Speech Communication*, vol. 94, pp. 62–74, 2017.

[16] P. Boersma, "Praat, a system for doing phonetics by computer," *Glot. International*, vol. 5, no. 9, pp. 341–345, 2001.

[17] A. Kelterer and B. Schuppler, "Turn-taking annotation for quantitative and qualitative analyses of conversation," *arXiv preprint arXiv:2504.09980*, 2025. [Online]. Available: `https://doi.org/10.48550/arXiv.2504.09980`.

[18] H. Buschmeier and M. Włodarczak, "TextGridTools: A TextGrid processing and analysis toolkit for Python," in *Tagungsband der 24. Konferenz zur Elektronischen Sprachsignalverarbeitung (ESSV 2013)*, 2013.

[19] J. Linke, S. Wepner, G. Kubin, and B. Schuppler, "Using Kaldi for Automatic Speech Recognition of Conversational Austrian German," in *arXiv*, 2023. arXiv: `2301.06475 [cs.CL]`. [Online]. Available: `https://doi.org/10.48550/arXiv.2301.06475`.

[20] A. Menrath, "Breath sounds and their relationship to turn-taking in conversational speech," M.S. thesis, Graz University of Technology, 2024.

[21] B. Schuppler, E. Berger, X. Kogler, and F. Pernkopf, "Homophone Disambiguation Profits from Durational Information," in *Interspeech 2022*, 2022, pp. 3198–3202.

[22] B. J. B. Schmitt, *AMFM_decompy Documentation 1.0.11*, `https://bjbschmitt.github.io/AMFM_decompy/`, Accessed: May 1, 2025, 2020. [Online]. Available: `https://bjbschmitt.github.io/AMFM_decompy/` (visited on 05/01/2025).

[23] S. A. Zahorian and H. Hu, "A spectral/temporal method for robust fundamental frequency tracking," *The Journal of the Acoustical Society of America*, vol. 123, no. 6, pp. 4559–4571, 2008, ISSN: 0001-4966. DOI: `10.1121/1.2916590`. eprint: `https://pubs.aip.org/asa/jasa/article-pdf/123/6/4559/14858368/4559\_1\_online.pdf`. [Online]. Available: `https://doi.org/10.1121/1.2916590`.

[24] M. Duarte, *Detecta: A Python module to detect events in data*, `https://doi.org/10.5281/zenodo.4598962`, version v0.0.5, detecta version 0.0.5, 2021. DOI: `10.5281/zenodo.4598962`. [Online]. Available: `https://doi.org/10.5281/zenodo.4598962`.

[25] M. Ali, *PyCaret: An open source, low-code machine learning library in Python*, PyCaret version 1.0.0, 2020. [Online]. Available: `https://www.pycaret.org`.

[26] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in *Advances in Neural Information Processing Systems 30*, Curran Associates, Inc., 2017, pp. 4765–4774. [Online]. Available: `http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf`.