Lexicon Meets Prosody: Classification of Cooperative & Competitive Overlaps

Tyrone White

***Abstract*—Speech overlaps in spontaneous conversation can be supportive backchannels or floor-grabbing interruptions. Although prosodic cues have been widely exploited, the incremental value of lexical information remains underexplored, especially in noisy, automatically transcribed meetings. I present a lightweight multimodal system trained on 1 006 automatically labeled overlap events from the AMI corpus. Adding sentence-level lexical embeddings to Wav2Vec audio features lifts macro F_1 by 0.14 absolute and boosts accuracy from 0.70 to 0.75, despite a mean word-error rate of 0.56 on competitive overlaps. A side experiment shows that flank context (-0.2 s left, $+0.3$ s right) unexpectedly degrades performance in the four-speaker setting, contradicting dyadic findings. The paper analyzes why lexical cues help and how multi-party dynamics blur prosodic context.**

Keywords: multi-party conversation, overlapping speech, multimodal classification

I. Introduction

In face-to-face dialog, listeners often speak before the current speaker yields the floor. Such *overlaps* vary in intent: a quick "yeah" may be cooperative, whereas cutting in with "but—" often signals competition. Distinguishing the two is critical for meeting assistants, call-center analytics, and conversational agents that must decide whether to cede or reclaim the turn.

Previous work has shown that local prosodic patterns (pitch rise, intensity bursts, and timing) carry discriminative power [1], [2]. Yet meetings also contain lexical markers of support (*yeah, right*) and disagreement (*no, but*) which a purely acoustic system cannot exploit. Two practical hurdles have limited lexical exploration so far: (i) high automatic speech recognition (ASR) error rates during overlaps, and (ii) the cost of manually labeling large overlap corpora.

In this study, I address both issues. I build a **weakly supervised** pipeline that combines simple heuristics with a large language model (LLM), DeepSeek-V3, to label 1 006 overlap clips (824 cooperative, 182 competitive) extracted from six scenario meetings of the AMI corpus [3]. Despite mean ASR word-error rates (WER) of 0.45 for cooperative and 0.56 for competitive overlaps, I show that adding sentence embeddings from the noisy transcripts to Wav2Vec audio features consistently improves accuracy and precision on the competitive class.

In the best setting (overlap-only segment, multimodal fusion), accuracy reaches 0.75 and macro-F_1 reaches 0.62, compared with 0.48 using text-only and 0.61 using audio-only. A bootstrap test across 10 000 resamples yields ΔF_1=0.016 with 95 % CI [–0.069, 0.095] and $p = 0.358$, indicating that the observed gain is not statistically significant at this scale.

The paper proceeds as follows: Section II surveys related work; Section III details data preparation and weak labeling; Sections IV–V present features and models; Section VI reports experiments; Section VII discusses findings; Section VIII concludes. The system instruction provided to the model is shown in Appendix A, and an example LLM prompt is included in Appendix B.

II. Related Work

A. Prosodic Approaches

Early efforts treated overlap intent as a prosody-only problem. Oertel *et al.* [1] analyzed 143 dyadic overlaps in the D64 corpus and found that acoustic and body-movement cues within a 0.2 s flank window best discriminated cooperative from competitive events, achieving 63 % accuracy with an SVM. Truong [2] used acoustic differences between overlapper and overlappee plus gaze for 355 AMI overlaps, reaching a 27.9 % equal-error rate after just 0.6 s of simultaneous speech.

B. Lexical and Multimodal Systems

Chowdhury *et al.* [4] introduced lexical n-grams and speaker-role features on 15 899 Italian call-center overlaps; an ensemble of acoustic and lexical models obtained an F_1 of 0.66. Galimzianov *et al.* [5] fine-tuned Conversational RuBERT on Russian customer-support dialogs and reported a macro F_1 of 0.84 on text alone, albeit with clean ASR and dyadic data.

Lebourdais *et al.* [6] reframed the task as detecting *interruptions* among seven overlap types in French broadcast speech; a lightweight WavLM classifier achieved 0.62–0.77 F_1 on the interruption class, depending on label consensus.

TABLE I
Prior work on overlap-intent classification.

Study	Corpus	Modality	Best Score
Oertel 12	D64	Prosody	0.63 (Acc)
Truong 13	AMI	Audio + gaze	0.28 (EER)
Chowdhury 15	Italian CC	Audio + Lexicon	0.66 (F_1*)
Lebourdais 24	French brdcst	Audio	0.77 (F_1*)
This work	AMI	Audio + Lexicon	0.62 (Macro F_1)

*On the competitive class.

C. Gap Addressed

Evidence for lexical benefits in noisy, multi-speaker meetings remains scarce (Table I). I fill this gap by quantifying the lexical boost on AMI data and by analyzing why very short flank context may hurt performance.

III. Data and Weak Labeling

A. Corpus

I use a subset of the AMI Meeting Corpus [3], consisting of six scenario-driven meetings (ES2002a/b, ES2005a/b, ES2012a/b) involving four participants in a collaborative design task. Headset audio and rich annotations make this set well-suited for fine-grained dialog analysis.

B. Overlap Extraction

Overlaps are defined as any temporal intersection between two speaker segments (from `segments.xml`) lasting at least 200 ms. I discard events with no lexical content (e.g., isolated coughs or laughs) and retain overlaps only if at least one channel contains words. From those, only overlaps involving laughter and lexical material are kept, since laughter can signal cooperation.

Each instance is associated with a context window: 4 s before and after the overlap. These 8 s are used for LLM labeling. Separately, I use only the 4 s pre-context for ASR transcription and text-based features, mimicking a real-time system where only past context is available.

C. Hybrid Labeling Pipeline

Heuristic stage. Obvious cooperative cases are caught via rules: e.g., the interrupter says only backchannels (*"yeah", "mm-hmm"*) or laughs while the main speaker continues.

LLM stage. Ambiguous cases are passed to DeepSeek-V3 with a prompt structured into *pre*, *overlap*, and *post* blocks, with speaker tags and timestamps (see Appendix B). The model chooses one of {cooperative, competitive, ambiguous}. Ambiguous outputs are dropped. If more than two participants speak, the LLM is prompted to label the overlap as competitive if any speaker behaves competitively; cooperative only if all appear supportive. The system prompt can be viewed in Appendix A.

D. Manual Review

To assess label reliability, I manually reviewed all 37 competitive instances in the test set and disagreed with only 4. This suggests that the LLM automated labeling yields competitive labels that are stable enough for modeling. I also verified many of the cooperative instances along the process and found them to be reliable.

E. Dataset Summary

Table II details the final label distribution.

TABLE II
Labeled overlap instances

Label	Count	Proportion
Cooperative	824	81.9 %
Competitive	182	18.1 %
Total	1 006	100 %

Audio for each overlap is stored as a `.flac` clip referenced in a manifest containing the clip path, timing, and label, enabling rapid batched feature extraction.

IV. Feature Extraction

A. Textual Features

For each overlap, I run Whisper-small to transcribe the 4 s pre-context plus the overlap segment. Transcripts are normalized (lowercased, punctuation removed, contractions expanded), and converted to 768-dimensional sentence embeddings using `all-mpnet-base-v2`. These embeddings serve as input to the text-only and fusion models.

Table III shows WER differences by class, computed against AMI reference transcriptions.

TABLE III
ASR word error rates (WER) by class

Label	Mean	Std Dev	Max
Cooperative	0.45	0.25	3.27
Competitive	0.56	0.28	3.00

B. Audio Features

From each overlap window, I extract 768-dimensional embeddings using `Wav2Vec2-base` with a mean-pooling operation. These are used as input to the audio-only model and concatenated with text embeddings in the fusion model.

Alternative window configurations (e.g., context padding) are explored in Section VI.

V. Models

All classifiers are lightweight feedforward neural networks predicting one of two classes: **cooperative** or **competitive**.

Text-only model. I use 768-dimensional MPNet sentence embeddings (via `sentence-transformers`) of the ASR-transcribed window. These are passed through a two-layer MLP with dimensions $768 \to 128 \to 2$, ReLU activation and dropout (rate 0.3). The final layer outputs raw logits to which you can apply softmax for evaluation.

Audio-only model. Wav2Vec2 embeddings extracted from the audio clip are passed to an identical MLP architecture.

Fusion model. Text and audio vectors are concatenated into a 1536-dimensional input, then passed through a deeper MLP (see Fig. 1): $1536 \to 256 \to 2$. This model jointly learns from lexical and prosodic cues.

All models are trained for 100 epochs using the Adam optimizer and cross-entropy loss.

VI. Experiments and Results

A. Evaluation Setup

I split the 1,006 labeled overlaps into a training and test set (80/20 stratified split). Evaluation is conducted on the 202-instance test set. I report accuracy, macro-averaged F_1, and per-class F_1 scores. I also include a bar chart (Fig. 2) to visualize precision, recall, and F_1 scores for the competitive class across all models.

B. Main Results

Table IV reports performance for each modality.

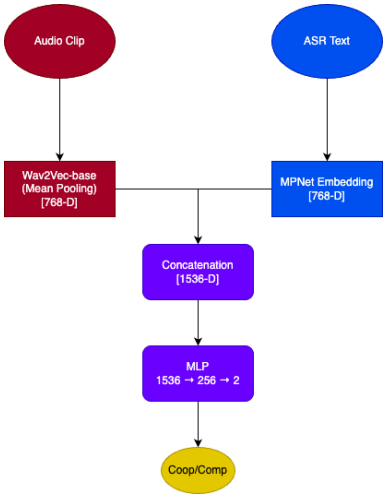

Fig. 1. Fusion model architecture: text and audio branches feed into a shared classifier.

TABLE IV
CLASSIFICATION RESULTS ON TEST SET (202 OVERLAPS)

Model	Acc	F_1 (macro)	F_1 (coop)	F_1 (comp)
Text-only	0.61	0.48	0.74	0.22
Audio-only	0.69	0.61	0.79	0.43
Fusion	0.75	0.62	0.84	0.41

C. Window Ablation

To test the effect of context length, I retrain the audio model on a short context window recommended by previous literature (0.2 s before to 0.3 s after overlap). Surprisingly, this significantly reduces accuracy (from 0.70 to 0.62) and competitive F_1 (from 0.43 to 0.33), suggesting that short flank audio may limit the model's ability to identify the overlap intent in multi-party settings.

D. Error Analysis

The text-only model frequently mislabels competitive overlaps that lack overt lexical interruptions. Audio features help capture prosodic intensity or floor grabs, improving competitive recall. The fusion model benefits from both inputs, but still struggles with subtle or ambiguous overlaps.

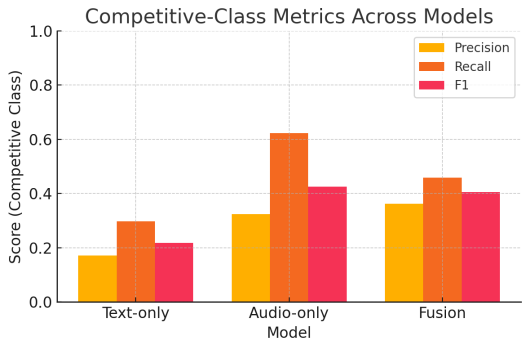

Fig. 2. Precision, recall, and F_1 for competitive overlaps across models.

VII. DISCUSSION

a) Lexical cues help—but trade-offs remain.: Figure 2 reveals that the audio-only model attains the highest *recall* for competitive overlaps (0.62) but at the cost of lower precision; the fusion model flips the emphasis, achieving the best precision (0.36) yet surrendering some recall (0.46). Their F_1 scores are therefore very close (0.43 vs. 0.41). Macro-averaged F_1 and overall accuracy favor the fusion model, suggesting that lexical information stabilizes predictions by boosting cooperative detection even if it occasionally reduces the high recall behavior of the audio branch on competitive overlaps. A paired bootstrap test, however, shows the macro F_1 gap ($\Delta = 0.016$) is not significant ($p = 0.36$). A larger evaluation corpus is required to verify whether the precision gain persists.

b) Competitive intent is inherently harder.: Competitive speech is more variable than cooperative back-channels. Speakers seize the floor with diverse lexical forms (questions, meta-comments, contentful disagreements) and a range of prosodic styles, from polite overlaps to aggressive cut-offs. Capturing this variety may require deeper dialog-level understanding, speaker-role modeling, and longer temporal context. In contrast, cooperative overlaps cluster around short acknowledgments and laughs, which the audio model already detects reliably.

c) Why short flank context hurts.: In dyadic dialogue the ± 0.2 s surrounding an overlap tends to carry useful prosodic cues from the same two speakers, so adding it boosts performance [1]. In AMI's four-speaker conversations that extra window is usually dominated by other voices or residual noise. These frames dilute the energy- and pitch patterns that distinguish a quick "mhm" from an attempted floor-grab, so the learned embedding drifts away from the actual overlap. In effect, the model trades signal for variance, leading to the observed 0.08 accuracy drop. Diarization or speaker-differenced prosodic features could reclaim the flank's potential without reintroducing noise.

d) Richer acoustic features.: The current system feeds a single 768-D Wav2Vec embedding per clip. Augmenting it with classic low-level descriptors pitch trajectories, intensity slopes, MFCC statistics) or explicit interrupter–overlappee prosodic contrasts could surface aggressive energy bursts that self-supervised vectors blur. Prior work found such hand-crafted cues informative [2]; combining them with Wav2Vec may improve precision without deeper networks.

e) Limitations.: I rely on six meetings recorded with headset microphones; results may not generalize to distant microphones or larger conferences. The weak labeling pipeline, while 89 % accurate by manual spot-check, may bias the classifier toward heuristic patterns. The dataset remains small by deep-learning standards; scaling up meetings, overlap instances, and model capacity is a direct avenue for improvement.

VIII. CONCLUSION

I presented a weakly supervised system that combines heuristics and an LLM to label 1 006 overlap events in AMI

meetings, then trains lightweight audio, text, and fusion models. Results suggest that lexical embeddings can boost overall performance over purely prosodic features, but gains remain modest and not yet statistically significant on the present test set. Competitive overlaps remain the bottleneck, pointing to the need for richer semantic modeling, speaker-aware features, and larger evaluation sets. Future work will scale both data and model size by collecting more meetings, incorporating stronger encoders, and exploring cross-modal attention, to evaluate how far simple capacity increases alone can push performance.

REFERENCES

[1] C. Oertel, T. Baumann, and G. Skantze, "Context cues for classification of competitive and collaborative overlaps," in *Proceedings of the 6th International Conference on Speech Prosody (Speech Prosody 2012)*. Shanghai, China: International Speech Communication Association (ISCA), 2012, pp. 22–25.

[2] 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.

[3] I. Mccowan, J. Carletta, W. Kraaij, S. Ashby, S. Bourban, M. Flynn, M. Guillemot, T. Hain, J. Kadlec, V. Karaiskos, M. Kronenthal, G. Lathoud, M. Lincoln, A. Lisowska Masson, W. Post, D. Reidsma, and P. Wellner, "The ami meeting corpus," *Int'l. Conf. on Methods and Techniques in Behavioral Research*, 01 2005.

[4] S. A. Chowdhury, M. Danieli, and G. Riccardi, "The role of speakers and context in classifying competition in overlapping speech," in *Interspeech 2015*, 2015, pp. 1844–1848.

[5] D. Galimzianov and V. Vyshegorodtsev, "Conversational rubert for detecting competitive interruptions in asr-transcribed dialogues," in *Computer Science and Information Technology*, ser. CSTY. Academy & Industry Research Collaboration Center, Jul. 2024, p. 67–75. [Online]. Available: http://dx.doi.org/10.5121/csit.2024.141306

[6] M. Lebourdais, M. Tahon, A. Laurent, and S. Meignier, "Automatic speech interruption detection: Analysis, corpus, and system," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, Eds. Torino, Italia: ELRA and ICCL, May 2024, pp. 1959–1968. [Online]. Available: https://aclanthology.org/2024.lrec-main.176/

APPENDIX

APPENDIX A. LLM SYSTEM INSTRUCTION

```
You are an expert annotator of conversational
overlaps in spontaneous multi-party speech. Your
task is to label each overlap as one of:

- cooperative: The overlap is a supportive or
affiliative action-such as a backchannel (\mm",
\mhm", \yeah", \uh-huh"), laughter, affirmation, or
short acknowledgment|that helps the main speaker
continue, signals agreement or understanding, or
indicates active listening. Cooperative overlaps
are not attempts to take the floor or redirect
the conversation. These include quick answers to
a question if it's clearly invited by the main
speaker and the first speaker does not resist.

- competitive: The overlap is an attempt to
interrupt, take the floor, or redirect the topic.
This includes interruptions, \grabbing the floor",
starting a new topic, or asking a new question
while the previous speaker is still talking.
If the overlapping speaker continues speaking after
the overlap and the first speaker falls silent,
this is likely competitive. Long or assertive
\yeah", \okay", \but...", or follow-up questions
can also be competitive if they shift the topic or
take the turn.

- ambiguous: The intent cannot be determined with
confidence, or the overlap does not clearly fit
either cooperative or competitive categories. Use
this only if the evidence is unclear or mixed.

You will be shown a snippet of a conversation
(chronological, with timing and speaker labels),
including a window of several seconds before,
during, and after the overlap. Use the full
context to judge intent.

Only respond with one of: cooperative, competitive,
ambiguous.

Labeling guidance:

- If the overlap is a short, anticipated answer
to a question and the first speaker finishes
naturally, this is cooperative.
- If more than two speakers overlap, use your
judgment; if any speaker is clearly trying to
take the floor, treat as competitive.
- If in doubt, or context is unclear,
respond ambiguous.

Respond ONLY with: cooperative, competitive, or
ambiguous.
```

APPENDIX B. LLM PROMPT EXAMPLE

```
You are classifying an overlap.
Return **only** one token: 'cooperative',
'competitive', or 'ambiguous'.

[PRE-CONTEXT]
1919.15s A: the
1919.28s A: the
1919.39s A: permanent
1920.74s A: mm
1921.16s A: battery
1921.75s A: ?
1921.83s B: Yeah
1922.79s B: ,
1922.79s B: think
[OVERLAP]
1922.93s B: that's
1922.97s D: Uh
1923.10s D: .
1923.10s D: That
1923.12s B: a
1923.17s B: good
1923.26s C: No
1923.29s B: idea
1923.32s C: .
[POST-CONTEXT]
1923.89s B: .
1924.19s D: sounds
1924.44s D: pretty
1924.63s D: good
1924.85s D: ,
1924.85s D: yeah
1925.04s D: .
1925.46s A: Is
1925.61s A: the
1925.88s A: uh
1926.13s A: you
1926.24s A: know
1926.41s A: ,
1926.41s A: we
1926.61s A: we
1926.62s C: Um
1926.79s C: .
1927.14s A: we
1927.26s A: are
1927.35s A: really
```