Contextual Speech Emotion Recognition with Large Language Models and ASR-Based Transcriptions

Enshi Zhang, Christian Poellabauer

Florida International University
{ezhan004,cpoellab}@fiu.edu

Abstract

Speech Emotion Recognition (SER) is the task of automatically identifying emotions expressed in spoken language. With the rise of large language models (LLMs), many studies have applied them to SER, but several key challenges remain. Current approaches often focus on isolated utterances, overlooking the rich contextual information present in conversations and the dynamic nature of emotions. Additionally, most methods rely on transcripts from a single Automatic Speech Recognition (ASR) model, neglecting the variability in word error rates (WER) across different ASR systems. Furthermore, the optimal length of conversational context and the impact of prompt structure on SER performance have not been sufficiently explored. To tackle these challenges, we design models using ASR transcripts from multiple sources as input data. In addition, we integrate custom prompts and different context window lengths. Empirical evaluations demonstrate that our method outperforms state-of-the-art techniques on the IEMOCAP and MELD datasets, highlighting the importance of utilizing conversational context and the diversity of ASR in SER tasks. All codes from our experiments are publicly available[1].

1 Introduction

Speech Emotion Recognition (SER) is an essential task in speech and natural language processing. The goal is to identify emotional states from spoken language [20, 21]. Advances in this field are valuable for the healthcare sector, where detecting emotional patterns in speech can assist in psychological evaluations and mental health diagnostics. Research has shown that individuals with mental health disorders often exhibit distinct emotional patterns in their speech [22, 9, 23], indicating the potential for speech-based diagnostic tools.

Although there is abundant publicly available speech data, much of it lacks emotional labels. The conventional method of labeling this data involves manual annotation by human annotators, a process fraught with challenges. Emotional interpretation is inherently subjective, and even with the involvement of domain experts or the consensus of multiple annotators to mitigate bias, the process remains laborious and costly [14, 8].

However, the recent advancement in LLMs' development offers a promising solution. These models, trained on extensive text corpora, can be harnessed for SER tasks with automatic speech recognition (ASR) [24] transcripts, potentially automating the emotional speech annotation process [20, 5]. Additionally, we can further explore how different factors, such as the length, structure, and accuracy of input texts and different prompts given to the LLM, affect the SER task results. For instance, human emotions develop within a conversation, making it essential to consider preceding dialogue

[1] https://github.com/coolsoda/SER_nips

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

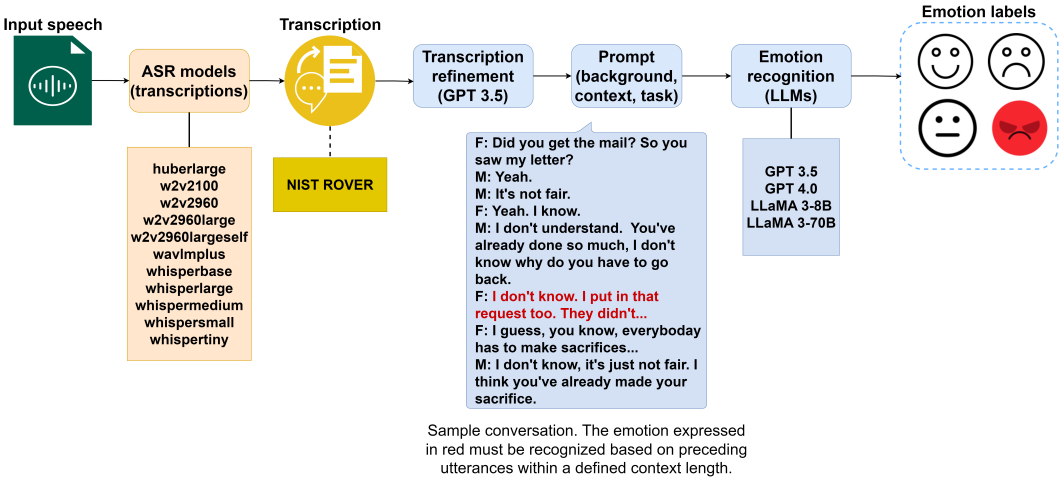

Figure 1: The workflow of our approach.

when predicting emotions in subsequent utterances. By incorporating conversational context, SER systems can better capture the complex emotional dynamics that unfold during interactions [12, 4, 7].

Our work is inspired by the IEEE SLT 2024 Generative Speech Error Correction (GenSEC) challenge [11][2]. We explore the capabilities of LLMs for emotion recognition, focusing solely on spoken text derived from ASR systems. We use the IEMOCAP [2] and MELD [19] datasets for our experiments. The key contributions of our study are as follows:

- We transcribed each audio utterance using eleven different ASR models to compare the word error rate (WER) in their transcripts. Our experiments uncovered that a lower WER does not always result in improved SER performance. We developed an algorithm that refines and selects the most suitable transcript from the eleven options, resulting in better SER performance than using the lowest WER transcript.
- In contrast to previous methods that process entire conversations sequentially, we focused on determining the optimal context length for LLM input. Our findings show that a longer context is not always more beneficial.
- We tested four LLMs as emotion annotators, demonstrating their efficiency in SER tasks. Additionally, we experimented with various prompt engineering techniques to determine the most effective approach for emotion recognition. Through extensive experiments on two SER datasets, our best-performing model surpassed existing state-of-the-art methods.

2 Methodology

2.1 Overview

We have outlined a three-step approach. First, we used models identified as 11 ASR, as illustrated in Figure 1, to transcribe emotional speech data. Second, we input the transcribed texts into both NIST Recogniser Output Voting Error Reduction (ROVER) system [6] and an LLM (GPT-3.5). ROVER is a post-processing algorithm that combines the outputs of multiple ASR systems to generate a new, improved (lower WER) output. The detailed ROVER algorithm is explained in Appendix Section C. In contrast, GPT-3.5 is instructed to select one of the 11 available transcripts, as noted in Algorithm 1 under section 2.3. Finally, we fed the selected text and the preceding sentences within the conversation, determined by a predefined context length, into different LLMs using a specially designed prompt to predict the final emotional label. Please note that ROVER's transcriptions have a significantly higher WER than most other ASR models, so we did not use them as the final input for our experiments.

[2] https://sites.google.com/view/gensec-challenge/home

	Neutral	Sad	Happy	Angry	Other	Overall
hubertlarge	†**0.20** ◇**0.21**	†**0.18** ◇0.20	†0.25 ◇0.26	†0.18 ◇0.18	†0.19 ◇0.18	†0.20 ◇ **0.21**
w2v2100	†0.41 ◇0.41	†0.34 ◇0.33	†0.43 ◇0.44	†0.37 ◇0.38	†0.36 ◇0.34	†0.38 ◇0.38
w2v2960	†0.35 ◇0.36	†0.28 ◇0.27	†0.33 ◇0.33	†0.26 ◇0.28	†0.28 ◇ 0.26	†0.29 ◇0.30
w2v2960large	†0.31 ◇0.32	†0.24 ◇0.23	†0.29 ◇0.28	†0.22 ◇0.21	†0.24 ◇0.23	†0.25 ◇0.27
w2v2960largeself	†0.23 ◇0.22	†**0.18** ◇**0.19**	†**0.22** ◇**0.23**	†**0.15** ◇**0.16**	†**0.18** ◇**0.16**	†**0.19** ◇0.22
wavlmplus	†0.45 ◇0.48	†0.39 ◇0.39	†0.43 ◇0.45	†0.32 ◇0.36	†0.37 ◇0.40	†0.38 ◇0.40
whisperbase	†0.36 ◇0.37	†0.40 ◇0.37	†0.48 ◇0.42	†0.34 ◇0.37	†0.38 ◇0.40	†0.39 ◇0.39
whisperlarge	†0.36 ◇0.36	†0.41 ◇0.42	†0.46 ◇0.48	†0.35 ◇0.36	†0.36 ◇0.39	†0.38 ◇0.41
whispermedium	†0.33 ◇0.34	†0.37 ◇0.35	†0.45 ◇0.40	†0.34 ◇0.31	†0.35 ◇0.31	†0.36 ◇0.34
whispersmall	†0.35 ◇0.35	†0.39 ◇0.38	†0.47 ◇0.44	†0.33 ◇0.29	†0.36 ◇0.30	†0.37 ◇0.35
whispertiny	†0.50 ◇0.42	†0.42 ◇0.35	†0.55 ◇0.49	†0.42 ◇0.39	†0.41 ◇0.33	†0.44 ◇0.40
ensemble	†0.42 ◇0.37	†0.55 ◇0.50	†0.25 ◇0.33	†0.31 ◇0.30	†0.28 ◇0.26	†0.32 ◇0.34
ROVER	†0.53◇0.51	†0.56 ◇0.52	†0.37 ◇0.37	†0.48 ◇0.45	†0.36 ◇0.33	†0.42 ◇0.40

Table 1: Word Error Rate (WER) analysis for all data points in the †IEMOCAP and ◇MELD datasets. The best results are shown in **bold**, and the second-best results are underlined.

2.2 ASR

To assess the impact of different transcriptions on SER performance, we utilized eleven ASR models to transcribe the emotional speech from both datasets. The transcriptions from the IEMOCAP dataset were organized in JSON format by the challenge organizers[3], where each key represents the utterance ID and the values correspond to the transcriptions generated by the different ASR models. Table 4 (see the Appendix, section A) provides an example of a data entry, demonstrating the substantial variation across the transcriptions. To investigate the impact of WER differences on SER performance, we referred to a previous study by [16] and presented the WER analysis in Table 1 for the eleven transcriptions across different emotional labels in both datasets. In order to address the issue of imbalanced label distribution, we adopted the strategy used in prior research, focusing on the four most commonly represented emotions: happy, sad, neutral, and anger [5, 3, 18, 20].

2.3 Transcription refinement

We observed significant variation in the transcriptions provided by different ASR models. To address this issue, we designed a straightforward algorithm (Algorithm 1), that, for each data entry, filtered out any transcriptions that are too short to provide enough information. If all transcriptions are short, we kept them all. Next, we refined the remaining transcriptions using ChatGPT 3.5. This refinement process selects the most coherent and logically structured transcription. If no clear choice is possible, we select the longest transcription, as it likely contains the most information. This process generates a new key for each data entry, which we call '**ensemble**'. We expect this approach to improve data quality for subsequent predictions.

Algorithm 1 Pseudo-code for ASR Transcriptions Refinement

Input: List of ASR transcriptions
Output: Refined transcription
1. Function GetFilteredTranscriptions(transcriptions, *min_length*)
  Keep transcriptions longer than *min_length*
  if none found, **return** all transcriptions
  return filtered transcriptions

2. Function RefineTranscriptionWithGPT(filtered_transcriptions)
  Construct prompt for GPT-3.5: "Choose the most comprehensive sentence. If unsure, select the longest one."
  return GPT-3.5 response

[3] https://github.com/YuanGongND/llm_speech_emotion_challenge?tab=readme-ov-file

Methods	F1 score				UA
	Neutral	Sadness	Happy	Anger	
whispertiny (WER †0.44 ◇0.40) + baseline_3 + GPT 3.5 (baseline)	†0.44 ◇0.53	†0.37 ◇0.20	†0.23 ◇0.32	†0.61 ◇0.55	†0.46 ◇0.48
w2v2960largeself (WER †0.19 ◇0.22) + baseline_3 + GPT 3.5	†0.44 ◇0.55	†0.51 ◇0.25	†0.42 ◇0.49	†0.67 ◇0.59	†0.54 ◇0.57
ensemble (WER †0.32 ◇0.34) + baseline_3 + GPT 3.5	†0.43 ◇0.55	†0.52 ◇0.27	†0.42 ◇0.51	†0.70 ◇0.59	†0.54 ◇0.58
ensemble + baseline_5 + GPT 3.5	†0.45 ◇0.56	†0.54 ◇0.27	†0.43 ◇0.52	†0.72 ◇0.61	†0.56 ◇0.58
ensemble + baseline_10 + GPT 3.5	†0.48 ◇0.57	†0.58 ◇0.30	†0.47 ◇0.55	†0.72 ◇0.62	†0.57 ◇0.60
ensemble + baseline_15 + GPT 3.5	†0.49 ◇0.57	†0.58 ◇0.29	†0.47 ◇0.57	†0.72 ◇0.60	†0.57 ◇0.59
ensemble + baseline_10 + GPT 4.0	†0.53 ◇0.61	†0.61 ◇0.32	†0.36 ◇0.58	†0.75 ◇0.63	†0.66 ◇0.63
ensemble + baseline_10 + LLaMA3-8B	†0.59 ◇0.62	†0.64 ◇0.35	†0.60 ◇0.59	†0.82 ◇0.64	†0.71 ◇0.65
ensemble + baseline_10 + LLaMA3-70B	†0.63 ◇0.66	†0.72 ◇0.36	†0.71 ◇0.59	†0.85 ◇0.66	†**0.75** ◇0.67
ensemble + expert_10 + LLaMA3-70B	†0.61 ◇0.64	†0.69 ◇0.34	†0.70 ◇0.58	†0.83 ◇0.64	†0.73 ◇0.65
ensemble + gambler_10 + LLaMA3-70B	†0.63 ◇0.74	†0.74 ◇0.36	†0.72 ◇0.60	†0.85 ◇0.67	†**0.75** ◇**0.69**
ensemble + CoT_10 + LLaMA3-70B	†0.62 ◇0.64	†0.68 ◇0.34	†0.66 ◇0.57	†0.80 ◇0.64	†0.72 ◇0.64
ensemble + CoT-fired_10 + LLaMA3-70B	†0.63 ◇0.64	†0.70 ◇0.34	†0.64 ◇0.56	†0.78 ◇0.62	†0.72 ◇0.64

Table 2: Summary of all experiments conducted. The corresponding WER, F1 score, and unweighted accuracy (UA) are reported for the †IEMOCAP and ◇MELD datasets. The best-performing method is highlighted in blue and bold.

2.4 Prompting

For this study, we utilized four large language models (LLMs): GPT-3.5, GPT-4.0[4], LLaMA 3-8B, and LLaMA 3-70B[5]. One challenge with ChatGPT is its sensitivity to prompt variations, which can lead to ambiguous or inaccurate results with even slight changes to the prompt. This poses a particular challenge when dealing with the IEMOCAP and MELD datasets, where human emotions in short dialogues can vary significantly [14]. Moreover, ChatGPT often raises concerns regarding reproducibility. To better assess performance, we used the two LLaMA models in parallel.

To identify the most effective prompting strategies for this task, we conducted comprehensive experiments inspired by previous studies [14, 1, 20]. Our exploration included analyzing the effect of mentioning subject-matter expertise in prompts, as well as the impact of irrelevant expertise or incentives to encourage correct responses. We also experimented with adding structured instructions to the prompts, including a step-by-step reasoning approach known as Chain-of-Thought (CoT). Another variation combined detailed instructions, step-by-step reasoning, and introduced penalties for incorrect answers. The prompt templates are provided in Table 5 under Appendix section B.

3 Datasets

IEMOCAP dataset comprises multi-modal recordings of human conversations from 10 subjects, with an equal split between males and females. It was created using transcripts from improvised video dialogues. The dataset includes 151 dyadic conversations, containing a total of 7,433 utterances, each categorized into one of six emotion labels: neutral, sad, angry, happy, frustrated, and excited. Consistent with prior studies [5, 3, 18, 20], we focused on the four most frequently used emotions—happy, angry, sad, and neutral—in our experiments.

MELD dataset consists of multi-party conversations extracted from the American TV show "Friends". It includes over 1,400 conversations, each containing between 1 and 33 utterances, resulting in a total of 13,708 utterances. Each utterance is annotated with one of seven emotion labels: anger, disgust, sadness, joy, neutral, surprise, and fear. For our experiments, we mapped the "joy" label to "happy" [4] and retained the labels for anger, sadness, and neutral.

4 Experiments and Results

4.1 Metrics

As our goal is to predict each utterance into one of the four emotional labels, regardless of its original ground truth, we followed established dominance protocols and used unweighted accuracy (UA)

[4] https://platform.openai.com/docs/api-reference/introduction

[5] https://docs.llama-api.com/quickstart

for SER. Additionally, F1 score were reported for each individual emotion class to provide a more detailed comparison against the baseline models.

4.2 Comparison results

Methods	IEMOCAP					MELD				
	Neutral	Sadness	Happy	Anger	UA	Neutral	Sadness	Happy	Anger	UA
MM-DFN [10]	0.66	0.80	0.44	0.68	0.68	0.77	0.24	0.54	0.48	0.61
COGMEN [13]	0.67	0.81	0.51	0.66	0.68	0.76	0.22	0.53	0.48	0.61
GCNet [17]	0.67	0.77	0.47	0.66	0.68	0.76	0.23	0.55	0.49	0.61
GraphCFC [15]	0.65	**0.85**	0.43	0.71	0.69	0.77	0.27	0.52	0.48	0.61
D^2GNN [4]	**0.68**	0.83	0.61	0.66	0.70	0.76	0.32	0.57	0.48	0.62
CoAttention-Acoustic [25]	–	–	–	–	0.72	–	–	–	–	–
LLM [5]	–	–	–	–	0.74	–	–	–	–	0.56
LLM-Acoustic [20]	–	–	–	–	**0.77**	–	–	—	–	–
Our approach	0.63	0.74	**0.72**	**0.85**	0.75	0.74	**0.36**	**0.60**	**0.67**	**0.69**

Table 3: Performance comparison on IEMOCAP and MELD datasets. We report F1 score for the emotional classes. The best performance is highlighted in bold.

In Table 2, we categorized all our experimental methods into five sections. In the first section, we conducted experiments using the ASR transcriptions with the lowest WER, specifically the w2v2960largeself model, while keeping all other variables aligned with the baseline method. The results showed significant improvements across all metrics. We also discovered that our 'ensemble' method achieved competitive, or even superior, performance despite having a higher WER. In the subsequent three sections, we continued using the 'ensemble' ASR transcriptions as our text input. We further investigated how context length (in the second section), different language models (in the third section), and varying prompts (in the fourth section) affect our SER tasks.

The results revealed several key insights. First, although the 'ensemble' method had a higher WER, as shown in Table 2, it slightly outperformed the w2v960largeself model in SER. This was particularly interesting for the 'sad' emotion, where the 'ensemble' method had the highest WER across both datasets but still delivered competitive results compared to other state-of-the-art methods. Second, increasing the context length from 5 to 10 significantly boosted performance, but further increasing it to 15 yielded little to no improvement, especially on the IEMOCAP dataset. Third, switching from GPT to LLaMA led to more robust performance, indicating that the size and architecture of LLMs play a crucial role in SER performance. Lastly, the 'gambler' prompt produced the best overall results, emphasizing the importance of prompt engineering in this task. These findings demonstrate the effectiveness of our approach, which combines ASR transcripts with LLMs and careful prompt design. However, from Table 3, our method performs well for predicting happiness and anger but struggles with 'neutral' emotions in both datasets. Moreover, the 'sadness' emotion posed challenges for most models, warranting further research.

5 Conclusion

In this study, we conducted a thorough investigation into the use of LLMs and ASR-transcribed text for SER. We focused on leveraging open APIs to reduce resource demands. Our findings revealed that a lower WER does not always result in better SER performance. We also confirmed that LLMs can effectively utilize extended conversational context to predict emotions more accurately than traditional approaches that handle each sentence independently. However, increasing the context length beyond a certain point does not constantly improve performance. Additionally, we underscored the pivotal role of prompt design in optimizing SER outcomes. Our analysis also demonstrated that LLMs excel in predicting emotions such as anger and happiness while distinguishing between emotions like happy and neutral remains more challenging.

Acknowledgment

We thank the anonymous reviewers for their invaluable feedback in enhancing our work and the organizers of the IEEE SLT 2024 GenSEC challenges for their support and encouragement.

References

[1] Mostafa M Amin and Björn W Schuller. On prompt sensitivity of chatgpt in affective computing. *arXiv preprint arXiv:2403.14006*, 2024.

[2] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359, 2008.

[3] Li-Wei Chen and Alexander Rudnicky. Exploring wav2vec 2.0 fine tuning for improved speech emotion recognition. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[4] Yijing Dai, Yingjian Li, Dongpeng Chen, Jinxing Li, and Guangming Lu. Multimodal decoupled distillation graph neural network for emotion recognition in conversation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2024.

[5] Tiantian Feng and Shrikanth Narayanan. Foundation model assisted automatic speech emotion recognition: Transcribing, annotating, and augmenting. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12116–12120. IEEE, 2024.

[6] Jonathan G Fiscus. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354. IEEE, 1997.

[7] Barbara Gendron and GaelGuibon GaelGuibon. Sec: Context-aware metric learning for efficient emotion recognition in conversation. In *Proceedings of the 14th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis*, pages 11–22, 2024.

[8] Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*, 2023.

[9] Yuan Gong and Christian Poellabauer. Topic modeling based multi-modal depression detection. In *Proceedings of the 7th annual workshop on Audio/Visual emotion challenge*, pages 69–76, 2017.

[10] Dou Hu, Xiaolong Hou, Lingwei Wei, Lianxin Jiang, and Yang Mo. Mm-dfn: Multimodal dynamic fusion network for emotion recognition in conversations. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7037–7041. IEEE, 2022.

[11] Chao-Han Huck Yang, Taejin Park, Yuan Gong, Yuanchao Li, Zhehuai Chen, Yen-Ting Lin, Chen Chen, Yuchen Hu, Kunal Dhawan, Piotr Żelasko, et al. Large language model based generative error correction: A challenge and baselines for speech recognition, speaker tagging, and emotion recognition. *arXiv e-prints*, pages arXiv–2409, 2024.

[12] Zhongquan Jian, Ante Wang, Jinsong Su, Junfeng Yao, Meihong Wang, and Qingqiang Wu. Emotrans: Emotional transition-based model for emotion recognition in conversation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5723–5733, 2024.

[13] Abhinav Joshi, Ashwani Bhat, Ayush Jain, Atin Singh, and Ashutosh Modi. Cogmen: Contextualized gnn based multimodal emotion recognition. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4148–4164, 2022.

[14] Siddique Latif, Muhammad Usama, Mohammad Ibrahim Malik, and Björn W Schuller. Can large language models aid in annotating speech emotional data? uncovering new frontiers. *arXiv preprint arXiv:2307.06090*, 2023.

[15] Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. Graphcfc: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia*, 26:77–89, 2023.

[16] Yuanchao Li, Peter Bell, and Catherine Lai. Speech emotion recognition with asr transcripts: A comprehensive study on word error rate and fusion techniques. *arXiv preprint arXiv:2406.08353*, 2024.

[17] Zheng Lian, Lan Chen, Licai Sun, Bin Liu, and Jianhua Tao. Gcnet: Graph completion network for incomplete multimodal learning in conversation. *IEEE Transactions on pattern analysis and machine intelligence*, 45(7):8419–8432, 2023.

[18] Leonardo Pepino, Pablo Riera, and Luciana Ferrer. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*, 2021.

[19] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

[20] Jennifer Santoso, Kenkichi Ishizuka, and Taiichi Hashimoto. Large language model-based emotional speech annotation using context and acoustic feature for speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11026–11030. IEEE, 2024.

[21] Vidhyasaharan Sethu, Julien Epps, and Eliathamby Ambikairajah. Speech based emotion recognition. In *Speech and Audio Processing for Coding, Enhancement and Recognition*, pages 197–228. Springer, 2014.

[22] Fuxiang Tao. *Speech-based automatic depression detection via biomarkers identification and artificial intelligence approaches*. PhD thesis, University of Glasgow, 2024.

[23] Daniela Teodorescu, Tiffany Cheng, Alona Fyshe, and Saif M Mohammad. Language and mental health: Measures of emotion dynamics from text as linguistic biosocial markers. *arXiv preprint arXiv:2310.17369*, 2023.

[24] Dong Yu and Lin Deng. *Automatic speech recognition*, volume 1. Springer, 2016.

[25] Heqing Zou, Yuke Si, Chen Chen, Deepu Rajan, and Eng Siong Chng. Speech emotion recognition with co-attention based multi-level acoustic information. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7367–7371. IEEE, 2022.

Key	Value
emotion	sad
id	Ses01F_script01_3_M023
Ground truth	'Yeah. I suppose I have been. But it's going from me.'
hubertlarge	'ya i suppose i have been bht's going from me'
w2v2100	'a i suppose i have been but's going from me'
w2v2960	'oh i suppose i have been let's going from me'
w2v2960large	'now i suppose i have been bat's going from me'
w2v2960largeself	'ar i suppose i have been but's going from me'
wavlmplus	'a i spose a habben was going for m'
whisperbase	'Yeah'
whisperlarge	'Yeah'
whispermedium	'Yeah'
whispersmall	'Yeah'
whispertiny	'Yeah'

Table 4: A sample data point from the IEMOCAP dataset, displaying all available transcriptions for a single utterance.

A Data Entry

Table 4 presents a sample dataset from the IEMOCAP dataset, which is partially provided by the organizers of the IEEE SLT GenSEC challenge. It is important to note that the WERs for the whisper models are typically higher than those of other models, as shown in Table 1. This is due to the fact that the whisper models were intentionally truncated to ensure they are not overly effective.

B Prompts

Table 5 presents five detailed prompts we used in our experiments. Each prompt includes background information, context (previous utterances up to a defined length), the current sentence (indicating the speaker and the sentence), and the task (consisting of prediction and rules).

C The ROVER Algorithm

Given multiple ASR transcriptions $\{\mathbf{T}_1, \mathbf{T}_2, \ldots, \mathbf{T}_n\}$ generated for the same audio input, the ROVER algorithm aims to produce a single combined transcription $\mathbf{T}_{\text{ROVER}}$ by aligning and voting on each word position. The steps are as follows:

- **Alignment:** Align each ASR output $\mathbf{T}_i$ at the word level using dynamic programming. Let each aligned sequence be represented as $\{\mathbf{w}_{i,1}, \mathbf{w}_{i,2}, \ldots, \mathbf{w}_{i,m}\}$, where $\mathbf{w}_{i,j}$ represents the word (or null if padded) from transcription i at position j.
- **Voting:** For each position j, consider the set of words $\{\mathbf{w}_{1,j}, \mathbf{w}_{2,j}, \ldots, \mathbf{w}_{n,j}\}$ across all ASR outputs. Use a majority vote to select the most frequent word $\mathbf{w}_j$ at each position j:

  $$\mathbf{w}_j = \text{argmax}_{\mathbf{w} \in \{\mathbf{w}_{1,j}, \mathbf{w}_{2,j}, \ldots, \mathbf{w}_{n,j}\}} \text{count}(\mathbf{w})$$

  If no majority exists, apply a tie-breaking rule, such as selecting a word from a higher-confidence ASR model.
- **Output Generation:** Construct the final combined transcription $\mathbf{T}_{\text{ROVER}} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_m\}$ by concatenating the selected words at each position j.

Name	Prompt Template
Baseline	Two speakers are talking. The conversation is {context}. Now speaker {current speaker} says: {current sentence}. Predict the emotion of this sentence from [happy, sad, neutral, angry], using the conversation context. Output only the label.
Expert	You are an expert emotion predictor. Analyze the conversation {context} and speaker {current speaker}. Now predict the emotion of the sentence {current sentence} from [happy, sad, neutral, angry], using the context. Output only the label.
Gambler	You are a gambler who earns money by predicting emotions correctly. The conversation is {context}. Now speaker {current speaker} says: {current sentence}. Predict the emotion of this sentence from [happy, sad, neutral, angry], using the context. Output only the label.
CoT	You are an expert emotion predictor. Analyze the conversation context {context} and speaker {current speaker}. Follow these steps: 1. Note emotional cues in the context. 2. Consider how they affect the current sentence. 3. Predict the emotion of {current sentence} from [happy, sad, neutral, angry]. Output only the label.
CoT-fired	You are an expert emotion predictor. Analyze the conversation {context}. Follow these steps: 1. Note emotional cues in the context. 2. Consider how they affect the current sentence. 3. Predict the emotion of {current sentence} from [happy, sad, neutral, angry]. Output only the label. If wrong, I will lose my job, so please try your best.

Table 5: The prompts designed for our experiments.

The resulting transcription $\mathbf{T}_{\text{ROVER}}$ is thus generated by consensus, potentially reducing errors found in individual ASR outputs.