



# Token-Level Logits Matter: A Closer Look at Speech Foundation Models for Ambiguous Emotion Recognition

Jule Valendo Halim, Siyi Wang, Hong Jia, Ting Dang

University of Melbourne, Australia

julevalendoh@student.unimelb.edu.au

## Abstract

Emotional intelligence in conversational AI is crucial across domains like human-computer interaction. While numerous models have been developed, they often overlook the complexity and ambiguity inherent in human emotions. In the era of large speech foundation models (SFMs), understanding their capability in recognizing ambiguous emotions is essential for the development of next-generation emotion-aware models. This study examines the effectiveness of SFMs in ambiguous emotion recognition. We designed prompts for ambiguous emotion prediction and introduced two novel approaches to infer ambiguous emotion distributions: one analysing generated text responses and the other examining the internal processing of SFMs through token-level logits. Our findings suggest that while SFMs may not consistently generate accurate text responses for ambiguous emotions, they can interpret such emotions at the token level based on prior knowledge, demonstrating robustness across different prompts.

**Index Terms:** emotion recognition, speech foundation models, multimodal large language models, affective computing

## 1. Introduction

Speech emotion recognition (SER) has experienced significant growth in areas such as mental health monitoring [1] and human-computer interaction (HCI) [2]. Recent advancements in large language models (LLMs) have led to the development of advanced speech foundation models (SFMs), which integrate speech inputs with LLMs. These models incorporate both verbal content (e.g., text transcription) and vocal nuances (e.g., sighs, laughter, tones) in an end-to-end manner, offering a more comprehensive understanding of speech content and a sophisticated framework for SER [3].

Despite the promises of these SFMs for emotion recognition, they primarily focus on identifying single emotion classes, such as categorizing a sentence as either happy or sad [4, 5]. However, human emotions are inherently multifaceted and ambiguous, influenced by cultural, contextual, and individual factors [6], which extend beyond simple classifications like happiness or sadness [7]. Yet, such emotion ambiguity is frequently neglected in automatic SER systems. This oversight mainly arises because the ground truth is set as a single label, based on majority voting from multiple annotations, which simplifies emotions into discrete categories [8]. This simplification fails to capture the ambiguity of human emotions, limiting SER systems ability to understand human emotions and, consequently, hindering applications in HCI and mental health assessments, where discerning subtle emotional states is essential.

While a few studies have examined ambiguous emotion recognition, they predominantly utilize traditional modelling

approaches such as Gaussian Mixture Regressions [9, 10], probabilistic networks [11], multi-task neural networks [12], Monte Carlo approach [13] or neural ordinary differential equations [14]. Only one recent study explored LLMs for ambiguous emotion recognition [15]. However, this study converts speech inputs into textual descriptions of features and uses text-based LLMs, an approach that does not process speech directly and fails to capture nuanced emotional information in speech.

This study is the first to explore ambiguous emotion recognition using end-to-end SFMs built on top of LLMs. The aim is to deepen our understanding of how these models recognize ambiguous emotions. We propose two approaches: the first analyses text responses directly to evaluate how SFMs “articulate” ambiguous emotions. The second approach examines how these models interpret ambiguous emotions by analyzing the intermediate representations from SFMs, thereby reflecting their intrinsic “conceptualization” of emotions. We propose a token-level analysis in which the logits of emotion-related tokens within SFMs are extracted and processed to represent ambiguous emotions.

By comparing the ambiguous emotion extracted from the model’s text responses (“articulation”) with the token-level representation of its internal processing (“conceptualization”), our findings reveal that *although SFMs may not consistently identify ambiguous emotions through articulation, they can more accurately process and recognize these ambiguous emotions at an token-level through posterior probability distributions*. Furthermore, the probability distributions are not sensitive to the prompts used, whereas the text responses are highly dependent on prompts. These insights highlight that pretrained SFMs possess the capabilities for recognizing ambiguous emotions, and our proposed token-level analysis provide an effective way to extract ambiguous emotion representation. This study opens up new avenues for enhancing chatbots or HCI applications through improved emotional intelligence in these systems.

## 2. Related Work

While the development of SER systems for single emotion classification has advanced for decades [16, 17, 18, 19], the progress in recognizing ambiguous emotions is still lagging behind. Emotion complexity and ambiguity were first recognized in [20], which recommended classifying emotions based on soft-labels rather than relying solely on categorical hard labels. Subsequent studies have proposed using multiple classifiers to mimic individuals and simulate the ambiguity in emotions [21], or treating ambiguous emotions as a completely different class, handling them as out-of-distribution data [22]. A few other studies have suggested modelling emotions as distributions [23, 10, 13]. However, these approaches are generally

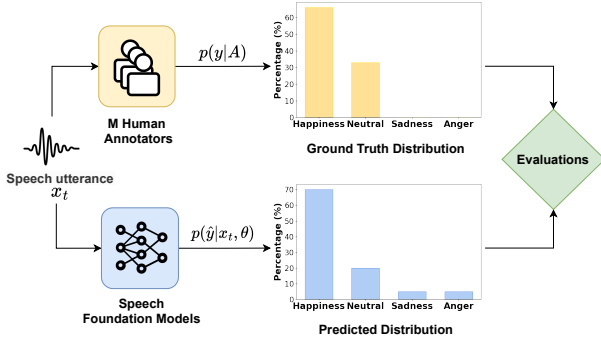


Figure 1: System overview. Speech utterances are processed by SFMs to generate emotion distributions, which are then compared with the ground truth inferred from  $M$  human annotators.

based on traditional modelling methods.

In the era of large-scale SFMs, employing a universal SFM trained on extensive speech data has opened up new possibilities for emotion understanding, moving away from reliance on task-specific small models. A recent study [24] explored the anthropomorphic capabilities of LLMs without fine-tuning, highlighting the capabilities of SFMs to capture emotional information from speech through expressive “delivery skills” inherently acquired during the training process. Another study [5] examined the capabilities of SFMs in recognizing single-label emotions with fine-tuning. However, these studies still overlook the ambiguity and complexity of human emotions. While [15] has shown that text-based LLMs can recognize ambiguous emotion to a certain extent, it does not explore the inherent capabilities of end-to-end SFMs processing speech directly.

### 3. Methodology

#### 3.1. Problem definition

As shown in Figure 1, given a speech utterance  $x_t$ , the objective is to infer the ambiguous emotion distribution  $p(\hat{y}|x_t, \theta)$  using a SFM parameterized by  $\theta$ . The predicted distribution encompasses  $N$  emotion classes, with each probability indicating the likelihood of a specific emotion, while the overall distribution reflects the inherent ambiguity in emotional expression. This predicted distribution is directly compared with the ground truth distribution  $p(y|A)$ , derived from the annotations  $A$  provided by  $M$  human annotators.

#### 3.2. System overview

As shown in Figure 2, we propose two approaches to predict the emotion distributions. The first approach focuses on text-level analysis that directly describe emotional ambiguity. The second approach conducts token-level analysis, extracting emotion distribution from the intermediate layers of SFMs. It examines how SFMs inherently process paralinguistic information in speech and the capabilities in emotion recognition they acquire during the pretraining paradigm. The prompt design will first be introduced, followed by two approaches for extracting emotion distributions.

#### 3.3. Prompt design

To guide SFMs in understanding ambiguous emotions, the prompt should explicitly direct them to generate appropriate responses. Three key components have been considered:

- *Emotional distribution prediction:* The prompt requests the likelihood of each emotion being represented in the speech

input, with probabilities expressed as percentages.

- *Logical reasoning:* The model is directed to use logical reasoning to determine the output percentages.

---

*Provide the likelihood (in percentages) that this audio represents each of the following emotions: anger, happiness, sadness, and neutral. Use logical reasoning to determine the percentages, but do not include this reasoning in your response.*

---

To ensure that the generated responses represent a valid distribution, where the probabilities in percentages sum to 100%, additional constraints on the outputs will be applied.

- *Output constraints:* If the generated responses for ambiguous emotion distributions fails to conform to a valid distribution, normalization is employed.

#### 3.4. Extraction of emotion distribution

##### 3.4.1. Approach 1: Text-level analysis

Given the speech utterances  $x_t$  and the prompt  $P$ , SFMs generate the text output describing the emotions (Figure 2). Each of the emotions is associated with corresponding percentages. A sample response given the prompt is as follows: Happiness: 0%, Neutral: 0%, Sadness: 35%, Anger: 65%, indicating that the SFM predominantly classifies the emotional content of the speech as anger, while also suggesting a probability of sadness. We convert the text responses into a discrete distribution and compare them with the ground truth distribution for evaluation.

##### 3.4.2. Approach 2: Token-level analysis

To further explore how SFMs inherently recognize ambiguous emotions, we focus on the token-level latent representations produced by the output layer of the SFMs, before they are converted into text output. It should be noted that SFM generated the output tokens in an autoregressive manner. For example, given the speech utterances  $x_t$  and the prompt  $P$ , the SFM generates the first token “ang”, followed by “er”, “6”, then “5”, and followed by “%”, until the complete text sequence was generated, represented by  $[\pi_1, \pi_2, \dots, \pi_J]$  where  $\pi_j$  represents the  $j^{th}$  output token. We extracted the logits  $z_j$  corresponding to each generated token  $\pi_j$  sequentially, and aggregated them across the entire  $J$  tokens to obtain the final logits representation to infer emotion distribution representations.

We propose a two-step process: first, extracting the logits  $z_j$  from emotion-related tokens in the vocabulary, and then converting these logits into the final emotion distributions  $\phi$ .

**Logit extraction from emotion-related tokens.** As shown in Figure 2, before generating each output token  $\pi_j$  (e.g., “Ang”), the inputs are first decoded to a latent vector that represents the vocabulary  $\mathcal{V}$ . This vector provides the probabilities for each token in the vocabulary, reflecting the likelihood of each being the next predicted token. We only extracted the logits corresponding to emotion-related tokens from the large vocabulary, e.g., [“ang” “er” “ne” “ut” “ral”, “sad” “ness”, “h” “app” iness], represented as  $z_j = [z_j^{ang}, z_j^{er}, z_j^{ne}, \dots, z_j^{iness}]$ .

As some emotion words are tokenized into multiple subword tokens, we average the logits across all subword tokens to obtain the logits for a single emotion word. For example, given “Anger” tokenized as “Ang” and “er”, we average their corresponding logits  $[z_j^{ang}, z_j^{er}]$ . This process is applied similarly for other emotions and is represented as:

$$z_j^{e_n} = \frac{1}{K} \sum_{k=1}^K z_{j,k}^{e_n} \quad (1)$$

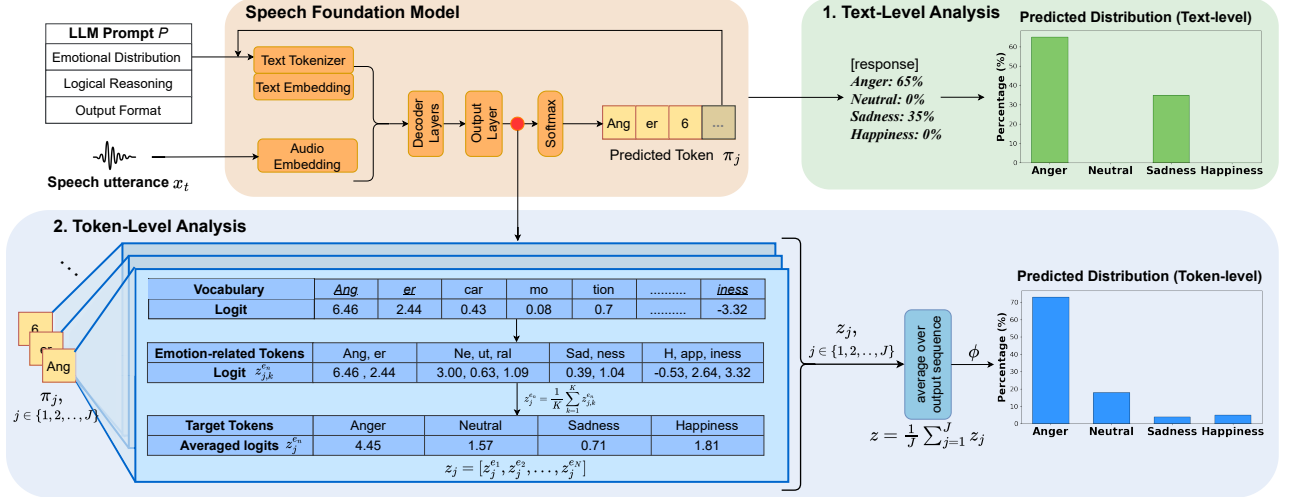


Figure 2: Framework for ambiguous emotion recognition using SFMs. By providing a prompt alongside speech, it enables the extraction of both the generated text and posterior probabilities at i) text-level and ii) token-level, respectively.

where  $K$  is the number of tokens for each emotion word  $e_n$ , and  $z_j^{e_n}$  is the averaged logits for the emotion  $e_n$ . This leads to a vector  $\mathbf{z}_j = [z_j^{e_1}, z_j^{e_2}, \dots, z_j^{e_N}]$  for each output token  $\pi_j$ .

Taking into account the entire output sequence, we average the logit representations  $\mathbf{z}_j$  across all output tokens  $\pi_j$ . This captures the emotion distribution across the whole output sequence as  $\mathbf{z} = \frac{1}{J} \sum_{j=1}^J \mathbf{z}_j = [z^{e_1}, z^{e_2}, \dots, z^{e_N}]$ .

**Conversion to Probabilities.** To make the logits a valid emotion distribution, the second step converts the logits into probabilities by normalizing over  $N$  emotions as follows:

$$\phi^{e_n} = \frac{z^{e_n}}{\sum_{n=1}^N z^{e_n}} \quad (2)$$

resulting in the final emotion distribution.

These normalized probabilities offer insights into the model’s inherent perception of emotional content and its internal decision-making process.

## 4. Experimental setup

### 4.1. Dataset

The Interactive Emotional Dyadic Motion Capture (IEMOCAP) database [25] is used, consisting of approximately 12 hours of audiovisual recordings of dyadic conversations between actors. Each recording is annotated by three annotators each. Specifically, we select utterances that only include four emotion classes: Happiness, Anger, Sadness, and Neutral, resulting in 4,373 speech files.

### 4.2. Implementation details

**Models.** We adopt LTU-AS [26] as the SFM which leverages LLaMA architecture. It is open-sourced, allowing flexibility for extracting intermediate representations. LTU-AS demonstrated strong performance in understanding emotion, due to its capacity to encode non-verbal cues such as sighs and laughter which are critical for understanding emotional content. LTU-AS integrates a Whisper encoder-decoder architecture [27] with LLaMA. The Whisper encoder converts both spoken and non-spoken audio into embeddings. Non-spoken audio embeddings are projected into a format compatible with LLaMA, while spoken audio is transcribed by the Whisper decoder into text.

This transcribed text is then converted into a distinct embedding, which is combined with the non-spoken audio embeddings and passed to LLaMA. Leveraging its pre-trained capabilities, LLaMA processes these embeddings, interprets emotional nuances, and generates a response that reflects emotional content from both verbal and non-verbal cues. Remote inference was performed through the HuggingFace Space platform, and local inference for posterior probabilities was conducted using two NVIDIA A100 80GB GPUs.

**Evaluation.** Two sets of evaluation metrics are used: ambiguity-based and accuracy-based. Ambiguity-based metrics aim to compare the predicted distributions and the ground truth distribution directly, including Bhattacharyya Distance (BD), Kullback-Leibler Divergence (KL), and the  $R^2$  Score. Additionally, we evaluate the performance of single label prediction to the majority vote of the annotations, by selecting the most likely emotion from the predicted distribution. While the model is not optimized for single emotion recognition, this evaluation offers a basic understanding of how it perceives the dominant emotion. Accuracy and F1-score are used.

## 5. Results

### 5.1. Performance on ambiguous SER

Table 1 presents the performance of ambiguous SER using *ambiguity-based metrics*, evaluating the entire distribution predictions. 2.1% of the outputs produced invalid distributions and were omitted from the evaluation. Our approach at the token level achieves the best KL divergence and BD, outperforming the baseline that converts speech to text descriptions and uses text-based LLMs.

More importantly, we found that the token-level analysis significantly outperforms the text-level analysis in recognizing ambiguous emotions, showing relative improvements of 51.71%, 7.84%, and 11.76% in terms of KL divergence, BD, and  $R^2$ , respectively. This highlights that *SFMs have the prior knowledge of ambiguity in emotional speech during the pre-training phase, but this ability does not fully translate into their text output*. The token-level analysis, therefore, provides an effective approach for inferring the nuances of ambiguous emotional expressions. While the baseline outperforms our ap-

Table 1: Performance for ambiguous SER. ( $\uparrow$ ) means higher values are better, and ( $\downarrow$ ) means lower values are better.

| Type     | Method         | KL ( $\downarrow$ ) | BD ( $\downarrow$ ) | $R^2$ ( $\uparrow$ ) |
|----------|----------------|---------------------|---------------------|----------------------|
| Baseline | Zero-Shot [15] | -                   | 0.51                | <b>0.51</b>          |
| Proposed | Text           | 2.05                | 0.51                | 0.34                 |
|          | Token-level    | <b>0.99</b>         | <b>0.47</b>         | 0.38                 |

Table 2: Impact of prompts for token-level analysis.

| Type        | Prompting | KL ( $\downarrow$ ) | BD ( $\downarrow$ ) | $R^2$ ( $\uparrow$ ) |
|-------------|-----------|---------------------|---------------------|----------------------|
| Token-level | Ambiguous | 0.99                | 0.47                | 0.38                 |
|             | Single    | <b>0.75</b>         | <b>0.31</b>         | <b>0.53</b>          |

proach in  $R^2$ , it is possibly due to their use of the advanced Gemini-1.5 model, whereas we use LLaMA-based (7B) models.

### 5.2. Impact of prompting at token level

To further investigate the robustness of token-level probabilities to variations in prompts for understanding emotional ambiguity, we employed prompts for single emotion recognition and compared them with prompts for ambiguous emotions. The single prompt we used is: “You are an expert in identifying emotions from speech. Predict the emotion of the audio from the choices [Happiness, Sadness, Neutral, Angry]. Respond with only one of the emotion labels.”. We also extracted the ambiguous emotion distributions at token-level for comparison.

As shown in Table 2, we found that: i) *token-level analysis remains robust against different prompting strategies for understanding ambiguous emotions. Even with single prompting, the model can grasp the ambiguity.* ii) Single-emotion prompting demonstrates superior performance in understanding ambiguous emotions, outperforming ambiguous prompting across all metrics. This further confirms that SFMs have an inherent understanding of emotional ambiguity, remaining robust despite variations in the prompts. iii) The best performance achieved with single prompting surpasses the baseline in Table 1, highlighting the effectiveness of our token-level analysis approach for ambiguous emotion prediction.

### 5.3. Token-level distribution extraction comparison

Instead of using the logits averaged across all output tokens  $\pi_j$  (section 3.4.2), we also examined the logits averaged only across the emotion word tokens, i.e., a subset of  $\pi_j$  that ignores the tokens related to percentages. This aims to identify the optimal distribution extraction from logits, as shown in Figure 3.

It is evident that incorporating logits from both the emotion-related text and the percentage outperforms using only the emotion-related text, suggesting that the logits for percentages also contain valuable emotional understanding. This is expected, as they reflect the model’s interpretation of the emotional categories. Moreover, SFMs operate in an autoregressive manner, where the percentage values capture meaningful information from the entire speech utterance as well as the previously generated text.

### 5.4. Single emotion prediction

To further understand how SFMs comprehend the dominant emotion, we infer the single emotion from our approaches and compared this with the ground truth majority vote emotion, as shown in Table 3. We first evaluated our approach using a single

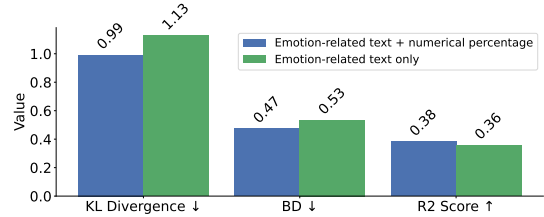


Figure 3: Performance comparison utilizing logits of i) both emotion-related text and numerical percentage and ii) only emotion-related text.

Table 3: Performance of single emotion recognition.

| Type       | Method                  | Accuracy     | F1-Score    |
|------------|-------------------------|--------------|-------------|
| Adaptation | Fine-Tuned LLM [28]     | <b>65.20</b> | -           |
|            | LLM Few-Shot [15]       | 58.75        | <b>0.59</b> |
| Zero-shot  | LLM Zero-Shot [15]      | 48.12        | 0.49        |
|            | LLM Zero-Shot [29]      | 51.40        | -           |
| Proposed   | Text (Single)           | <u>58.06</u> | <u>0.39</u> |
|            | Text (Ambiguous)        | 35.83        | 0.06        |
|            | Token-Level (Ambiguous) | 52.64        | 0.31        |
|            | Token-Level (Single)    | 55.90        | 0.30        |

prompt for single-emotion prediction and analyzed the text outputs for emotion recognition. This method achieved a superior accuracy of 58.06%, outperforming other zero-shot learning approaches. These results highlight the advantages of leveraging speech input rather than relying solely on text input, as it enables more comprehensive processing of emotional traits from speech. It shows inferior performance compared to the LLM baselines with fine-tuning or few-shot learning, which is expected due to the additional training or the few-shot examples.

Additionally, we inferred the most likely class from the predicted ambiguous distributions and compared it to the majority vote using both text-level and token-level analysis. It is important to note that the prompt for ambiguous emotions and the token-level analysis were not optimized for single emotion prediction. Nevertheless, the token-level analysis still achieved comparable or superior accuracy to zero-shot baselines. The lower F1-score is possible due to the inherent bias in SFMs towards the majority class. In summary, our approach not only captures the full emotion spectrum but also identifies the dominant emotion to a certain extent.

## 6. Conclusion

Our study investigated the extent to which pretrained SFMs can interpret ambiguity in SER based on their prior knowledge and introduced two approaches to infer ambiguous emotion distributions at both the text and token levels. Our findings suggest that SFMs can recognize the nuanced ambiguity present in emotional speech due to its prior knowledge acquired during the pretraining phase, while this is not fully translated in their text-based outputs. The proposed token-level analysis offers an effective method for inferring both ambiguous emotion distributions and single emotions, outperforming state-of-the-art zero-shot baselines. These results highlight the potential of SFMs in applications where recognizing complex and ambiguous emotional states is crucial, such as mental health monitoring, and offer an innovative method to infer emotions from SFMs.

## 7. References

- [1] N. Elsayed, Z. ElSayed, N. Asadizanjani, M. Ozer, A. Abdelgawad, and M. Bayoumi, "Speech emotion recognition using supervised deep recurrent system for mental health monitoring," 2024.
- [2] S. Ramakrishnan and I. M. El Emary, "Speech emotion recognition approaches in human computer interaction," *Telecommunication Systems*, vol. 52, pp. 1467–1478, 2013.
- [3] Z. Wu, Z. Gong, L. Ai, P. Shi, K. Donbekci, and J. Hirschberg, "Beyond silent letters: Amplifying llms in emotion recognition with vocal nuances," 2024.
- [4] M. Niu, M. Jaiswal, and E. M. Provost, "From text to emotion: Unveiling the emotion annotation capabilities of llms," *INTERSPEECH*, 2024.
- [5] T. Feng and S. 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)*. IEEE, 2024, pp. 12 116–12 120.
- [6] E. Mower, A. Metallinou, C. M. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *Proceedings of the International Conference on Affective Computing and Intelligent Interaction*, 2009, pp. 1–8.
- [7] J. Wu, T. Dang, V. Sethu, and E. Ambikairajah, "Emotion recognition systems must embrace ambiguity," 2024.
- [8] W. Wu, B. Li, C. Zhang, C.-C. Chiu, Q. Li, J. Bai, T. N. Sainath, and P. C. Woodland, "Handling ambiguity in emotion: From out-of-domain detection to distribution estimation," 2024. [Online]. Available: <https://arxiv.org/abs/2402.12862>
- [9] T. Dang, V. Sethu, J. Epps, and E. Ambikairajah, "An investigation of emotion prediction uncertainty using gaussian mixture regression," in *INTERSPEECH*, 2017, pp. 1248–1252.
- [10] T. Dang, V. Sethu, and E. Ambikairajah, "Dynamic multi-rater gaussian mixture regression incorporating temporal dependencies of emotion uncertainty using kalman filters," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4929–4933.
- [11] M. N. Mohanty and H. K. Palo, "Child emotion recognition using probabilistic neural network with effective features," *Measurement*, vol. 152, p. 107369, 2020.
- [12] Y. Zhang, J. Fu, D. She, Y. Zhang, S. Wang, and J. Yang, "Text emotion distribution learning via multi-task convolutional neural network," in *IJCAI*, 2018, pp. 4595–4601.
- [13] J. Wu, T. Dang, V. Sethu, and E. Ambikairajah, "A novel sequential monte carlo framework for predicting ambiguous emotion states," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 8567–8571.
- [14] J. Wu, T. Dang, V. Sethu, and E. Ambikairajah, "Dual-constrained dynamical neural odes for ambiguity-aware continuous emotion prediction," in *Proc. Interspeech 2024*, 2024, pp. 3185–3189.
- [15] X. Hong, Y. Gong, V. Sethu, and T. Dang, "Aer-llm: Ambiguity-aware emotion recognition leveraging large language models," in *ICASSP*, 2024.
- [16] A. A. Abdelhamid, E.-S. M. El-Kenawy, B. Alotaibi, G. M. Amer, M. Y. Abdelkader, A. Ibrahim, and M. M. Eid, "Robust speech emotion recognition using cnn+ lstm based on stochastic fractal search optimization algorithm," *Ieee Access*, vol. 10, pp. 49 265–49 284, 2022.
- [17] Z. Yang, Z. Li, S. Zhou, L. Zhang, and S. Serikawa, "Speech emotion recognition based on multi-feature speed rate and lstm," *Neurocomputing*, vol. 601, p. 128177, 2024.
- [18] C. Fang, Y. Jin, G. Chen, Y. Zhang, S. Li, Y. Ma, and Y. Xie, "Multimodal speech emotion recognition based on large language model," *IEICE TRANSACTIONS on Information and Systems*, vol. 107, no. 11, pp. 1463–1467, 2024.
- [19] C. O. Kumar, N. Gowtham, M. Zakariah, and A. Almazayad, "Multimodal emotion recognition using feature fusion: An llm-based approach," *IEEE Access*, 2024.
- [20] E. Mower, A. Metallinou, C.-C. Lee, A. Kazemzadeh, C. Busso, S. Lee, and S. Narayanan, "Interpreting ambiguous emotional expressions," in *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*. IEEE, 2009, pp. 1–8.
- [21] Y. Zhou, X. Liang, Y. Gu, Y. Yin, and L. Yao, "Multi-classifier interactive learning for ambiguous speech emotion recognition," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 30, pp. 695–705, 2022.
- [22] W. Wu, B. Li, C. Zhang, C.-C. Chiu, Q. Li, J. Bai, T. N. Sainath, and P. C. Woodland, "Handling ambiguity in emotion: From out-of-domain detection to distribution estimation," *arXiv preprint arXiv:2402.12862*, 2024.
- [23] M. Atcheson, V. Sethu, and J. Epps, "Using gaussian processes with lstm neural networks to predict continuous-time, dimensional emotion in ambiguous speech," in *2019 8th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2019, pp. 718–724.
- [24] J.-t. Huang, M. H. Lam, E. J. Li, S. Ren, W. Wang, W. Jiao, Z. Tu, and M. R. Lyu, "Emotionally numb or empathetic? evaluating how llms feel using emotionbench," 2023.
- [25] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.
- [26] Y. Gong, A. H. Liu, H. Luo, L. Karlinsky, and J. Glass, "Joint audio and speech understanding," in *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 2023, pp. 1–8.
- [27] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," *PMLR*, pp. 28 492–28 518, 2023.
- [28] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," 2024.
- [29] S. Feng, G. Sun, N. Lubis, W. Wu, C. Zhang, and M. Gašić, "Affect recognition in conversations using large language models," *arXiv preprint arXiv:2309.12881*, 2023.