Multiclass Fairness Analysis of Qwen2-Audio in Speech Emotion Recognition

Federico D'Asaro[1,2], Juan José Márquez Villacís[2], Andrea Bottino[1], and Giuseppe Rizzo[1,2]
[1]Dipartimento di Automatica e Informatica, Politecnico di Torino, Turin, Italy
[2]AI, Data & Space (ADS), LINKS Foundation, Turin, Italy

***Abstract*—Recognizing emotions from speech using advanced audio-based Large Language Models (ALMs) has shown promise due to their zero-shot capabilities, with applications in healthcare, education, and human–computer interaction. However, fairness evaluation of ALMs across sensitive groups such as gender, age, and ethnicity remains largely unexplored. In this work, we evaluate Qwen2-Audio, the strongest open-source ALM we benchmark for Speech Emotion Recognition (SER), using metrics including Statistical Parity, Equal Opportunity, and Overall Accuracy Equality, while accounting for the multiclass nature of emotions. Our results indicate that Qwen2-Audio exhibits unfairness across datasets and sensitive attributes in terms of OAE, emphasizing the need for fairness-aware SER models. All the code associated with this paper is available at: https://github.com/links-ads/FairALM-Emotion-Recognition.**

***Index Terms*—Audio Language Models, Speech Emotion Recognition, Fair Machine Learning.**

I. Introduction

Speech Emotion Recognition (SER) extracts emotional information from spoken audio and plays an essential role in various Human–AI interaction technologies employed in sectors like healthcare, security and education [1]. Recently, Audio-Language Models (ALMs) [2] have emerged as multimodal extensions of Large Language Models, capable of ingesting input audio and text queries to accomplish speech-related tasks such as SER in a zero-shot fashion.

Current studies employing ALMs move in several directions. Some explore their capability to understand the paralinguistic aspects of speech [3]. Others evaluate how prompts incorporating emotion-specific knowledge from acoustics, linguistics, and psychology can benefit the task [4]. Additional work proposes emotion-specific ALMs specialized through instruction tuning [5].

However, little attention has been given to assessing the fairness of ALMs in emotion prediction. Fairness refers to evaluating whether automated decision-making systems behave equitably toward individuals [6]. A common approach assesses fairness by exploiting sensitive attributes such as gender in a binary classification setting [7], [8]. While available SER datasets often include sensitive attributes such as gender, age, and ethnicity, fairness assessment presents two main challenges: (i) Sensitive attributes may have more than two categories (e.g. age, ethnicity), known as the *multigroup* setting, and (ii) SER is inherently a *multiclass* classification task, in contrast to the binary setting where most standard fairness metrics are defined.

In this work, we take a first step toward assessing the fairness of ALMs in SER. We first evaluate distinct open-source ALMs [2], [9]–[11] on SER and find that Qwen2-Audio achieves the best performance. Motivated by recent efforts to extend fairness assessment to multigroup–multiclass scenarios [12], we perform such an evaluation for SER using Qwen2-Audio on benchmarks including CREMA-D, IEMOCAP, EmoV-DB, RAVDESS, and MELD (see Table I).

Our aim is diagnostic: we provide the first fairness audit of ALMs in SER and a reusable evaluation setup, rather than a mitigation method, which we leave to future work.

The main contributions of this work are:

- We analyze both the performance and fairness of the Qwen2-Audio ALM in SER, also examining the impact of generative temperature.
- We adopt a recent approach for multiclass fairness assessment [12] and apply it to the SER task.
- We show that unfairness in Qwen2-Audio for SER manifests in terms of Overall Accuracy Equality (OAE).

II. Related Work

A. Speech Emotion Recognition

Early SER approaches primarily relied on handcrafted acoustic features, including prosodic, spectral, and voice quality cues [1], combined with traditional machine learning classifiers such as KNN, SVM, and Naïve Bayes [13]. The field progressed with the adoption of deep learning, from Convolutional Neural Networks to Transformer-based models [14]. The emergence of Large Speech Models, pretrained on massive audio corpora using self-supervised objectives [15] or weak audio–text alignment signals [16], further advanced SER by providing powerful universal audio representations; many works build upon these models by training small classifiers atop frozen LSM features [17].

More recently, Audio-Language Models (ALMs) have emerged as LLM-based architectures capable of processing audio inputs and performing a broad range of zero-shot audio tasks, including emotion recognition. Among open-source ALMs [2], [9]–[11], we select Qwen2-Audio as the best-performing model. Being a generative model, we evaluate it across various temperature settings [18] to assess whether it behaves fairly in SER.

B. Fairness in Machine Learning

As machine learning and automated decision-making systems became pervasive throughout the 2010s, concerns about their equitable behavior emerged as a major research focus [6]. A core concept in this area is the *sensitive attribute*, which denotes aspects of human-related data that carry social or ethical significance—such as gender, ethnicity, or age—and whose misuse can lead to harmful or discriminatory outcomes [19]. Evaluating fairness remains a complex problem, and a wide range of metrics has been proposed. These metrics are commonly grouped into two paradigms: *Group Fairness*, which examines how predictive performance differs across subsets of the population defined by a protected characteristic [7], [8], [20], and *Individual Fairness*, which instead considers whether similar individuals receive similar model outputs [21].

In this work, we focus on Group Fairness in SER, leveraging sensitive attributes commonly available in datasets. Most fairness metrics are designed for binary tasks, and few studies consider multiclass settings. Building on recent advances in this area, we adopt multiclass fairness metrics from [12] to the SER context, representing the first exploration in this direction. Such a study is useful because it allows us to quantify fairness across multiple aspects, including group-level prediction balance, equality of correct predictions, and overall performance consistency, offering a more comprehensive view of potential biases.

III. Group Fairness Metrics

In this work, we focus on *Group Fairness* metrics, which compare the outcomes of a classification algorithm across two or more groups. Here, fairness concerns the model's emotion predictions: after mapping each generated response to a class label (Section IV), we require parity of predictions (SP), error rates (EO), and accuracy (OAE) across groups, not the surface form of the generated text. According to recent reviews on machine learning fairness, *Statistical Parity*, *Equal Opportunity*, and *Overall Accuracy Equality* are well-defined and widely used fairness metrics [22].

Notation. Let $Y \in \{0, 1\}$ denote the true labels and $\hat{Y} \in \{0, 1\}$ the corresponding predicted labels. The sensitive attribute A takes values in a finite set $\{a_1, \ldots, a_{|A|}\}$. For each group a_i, we write $\Pr(\cdot \mid a_i)$ to denote probabilities conditioned on membership in group a_i.

In the following, we define the metrics in the binary classification setting for a given pair of groups (a_i, a_j).

Statistical Parity (SP). Statistical Parity [7] requires each group to have the same probability of being classified as positive. It is defined as the absolute difference in positive prediction rates between two groups:

$$U_{\mathrm{SP}} = \left| \Pr(\hat{Y} = 1 \mid a_i) - \Pr(\hat{Y} = 1 \mid a_j) \right|. \tag{1}$$

Equal Opportunity (EO). Equal Opportunity [8] requires equality of True Positive Rates between groups:

$$U_{\mathrm{EO}} = \left| \Pr(\hat{Y} = 1 \mid Y = 1, a_i) - \Pr(\hat{Y} = 1 \mid Y = 1, a_j) \right|. \tag{2}$$

TABLE I: Selected Speech Emotion Recognition Datasets

Dataset	# Speakers	Sensitive Attributes	# Emotion
CREMA-D	96 (48M, 48F)	Gender, Age, Ethnicity	6
IEMOCAP	10 (5M, 5F)	Gender	4
EmoV-DB	4 speakers	Gender	5
RAVDESS	24 (12M, 12F)	Gender	8
MELD	6 (3M, 3F)	Gender	7

Overall Accuracy Equality (OAE). Overall Accuracy Equality [20] requires accuracy to be equal across groups:

$$U_{\mathrm{OAE}} = \left| \Pr(\hat{Y} = Y \mid a_i) - \Pr(\hat{Y} = Y \mid a_j) \right|, \tag{3}$$

Multigroup Unfairness. All the definitions above naturally apply to the case $|A| = 2$, where the sensitive attribute A consists of two groups. They can also be extended to the non-binary case, i.e., $|A| = m > 2$, by averaging the fairness indices over all pairs of groups. Formally,

$$U(a_1, \ldots, a_m) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} U(a_i, a_j). \tag{4}$$

The unfairness measures lie in $[0, 1]$ (0 under perfect fairness) and are reported as percentages in $[0, 100]$.

Multiclass and Multigroup Unfairness. Since most Emotion Recognition datasets involve multiclass classification with labels such as *Happy*, *Surprised*, and *Sad* (see Table I), the metrics defined above do not directly apply to this setting and must be further generalized to the non-binary case.

Following [12], we address the multiclass problem by reducing it to a set of binary subproblems. Let $Y \in \{0, 1, \ldots, K\}$ be the multiclass label. For each class c, we define a corresponding binary task:

$$Y_c = \begin{cases} 1, & Y = c, \\ 0, & Y \neq c, \end{cases} \tag{5}$$

allowing any binary group-fairness metric to be computed classwise. Let $U_c(a_i, a_j)$ denote the unfairness measure (e.g., Statistical Parity) for the binary task Y_c. The classwise unfairness $U_c(a_1, \ldots, a_m)$ is then obtained by applying the pairwise averaging of Eq. 4, yielding a score for each of the $K + 1$ classes.

An overall multiclass unfairness measure can then be obtained by averaging the classwise values:

$$U_{\mathrm{multiclass}} = \frac{1}{K+1} \sum_{c=0}^{K} U_c. \tag{6}$$

OAE is reported only at the group level, not classwise.

IV. Emotion Recognition Using ALM

In this section, we describe the Speech Emotion Recognition pipeline using Audio-Language Models (ALMs), illustrated in Figure 1. We outline the ALM architecture, the Task Prompt, and the adopted Levenshtein-based post-processing method.

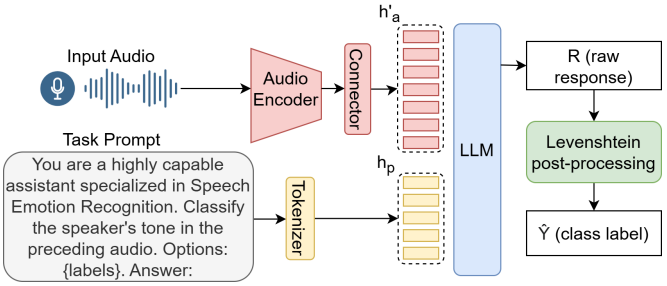


Fig. 1: Our Emotion Recognition pipeline exploiting ALMs.

Architecture. An ALM can be abstracted into three modules: (i) an *audio encoder*, (ii) a *connector*, and (iii) an *LLM*.

The audio encoder (e.g., Whisper [16]) extracts features from the input audio x, producing a sequence of audio tokens $h_A \in \mathbb{R}^{n_A \times d_A}$. The connector projects and downsamples these tokens into the LLM embedding space, yielding $h'_A \in \mathbb{R}^{n'_A \times d_{\text{LLM}}}$ with $n'_A < n_A$. The audio tokens are concatenated with the task-prompt tokens $h_p \in \mathbb{R}^{n_p \times d_{\text{LLM}}}$ and passed to the LLM, which generates a textual response:

$$R = \text{LLM}(\,[h'_A; h_p]\,)\,. \tag{7}$$

Task Prompt. The task prompt instructs the LLM on the emotion classification task. Since the task is categorical, the prompt explicitly includes the candidate emotion labels of the dataset on which the model is evaluated.

Temperature. Text generation proceeds token-by-token: from logits $[l_1, \ldots, l_v]$, where v is the vocabulary size, a probability distribution $[p_1, \ldots, p_v]$ is obtained. A token is then sampled according to these probabilities. The distribution is controlled by the temperature parameter τ: high values yield a flatter (more uniform) distribution, while low values sharpen it, making the sampling more deterministic. Following prior work [18], we investigate the effect of temperature on SER performance and fairness.

Post-Processing. Due to the generative nature of ALMs, responses R may differ from ground-truth labels Y in form or length (e.g., adjectives instead of nouns). Following [23], we apply a Levenshtein-based post-processing step.

If R is not an exact match, it is normalized and tokenized. We compute the Levenshtein similarity between each token and each target label:

$$\text{LevRatio}(\text{label}, \text{word}) = 1 - \frac{\text{LevDist}(\text{label}, \text{word})}{\text{len}(\text{label}) + \text{len}(\text{word})}. \tag{8}$$

Scores below 0.57 are discarded, and the label with the highest total similarity is selected [23].

V. Experiments

Datasets. We evaluate on five widely used SER datasets (Table I): *CREMA-D* [24], whose 96 actors span diverse ethnic, racial, and age groups; *IEMOCAP* [25]; *EmoV-DB* [26]; *RAVDESS* [27]; and *MELD* [28], a conversational corpus from *Friends* dialogues. For MELD we limit the analysis to the six main characters, accounting for 2,156 of 2,610 test samples, excluding others due to missing actor information.

TABLE II: Macro-F1 (%) of fine-tuned EmoBox baselines (top) and zero-shot ALMs (bottom) on CREMA-D, IEMOCAP, EmoV-DB, RAVDESS, and MELD. The best zero-shot ALM per dataset is underlined; baselines are fine-tuned per dataset and shown for reference only.

Model	CREMA-D	IEMOCAP	EmoV-DB	RAVDESS	MELD
HuBERT large	73.73	67.24	99.37	69.54	24.99
WavLM large	74.39	69.29	99.45	71.42	29.11
data2vec large	63.48	51.71	94.58	58.74	24.10
data2vec 2.0 large	69.25	56.70	98.19	70.94	27.35
Whisper large v3	76.60	73.11	99.34	75.19	32.95
Qwen2-Audio	<u>76.56</u> (± 2.21)	<u>70.42</u> (± 6.75)	69.64 (± 1.54)	<u>67.48</u> (± 5.34)	<u>28.11</u> (± 2.07)
Audio-Flamingo-3	61.00 (± 2.80)	46.95 (± 2.38)	<u>83.85</u> (± 0.81)	25.94 (± 3.34)	20.03 (± 1.22)
Salmonn-7B	7.79 (± 3.07)	31.33 (± 6.12)	15.69 (± 6.24)	3.99 (± 1.98)	7.23 (± 1.73)
Voxtral-Mini	6.02 (± 0.64)	30.21 (± 2.42)	10.79 (± 0.97)	4.71 (± 1.78)	24.75 (± 3.31)

TABLE III: F1 results of Qwen2-Audio with and without Levenshtein post-processing.

Model	CREMA-D	IEMOCAP	EmoV-DB	MELD
w/ Lev.	<u>76.56</u> (± 2.21)	<u>70.42</u> (± 6.75)	<u>69.64</u> (± 1.54)	<u>28.11</u> (± 2.07)
w/o Lev	74.93 (± 2.98)	69.60 (± 7.65)	67.53 (± 1.05)	26.84 (± 2.02)

Metrics. In our evaluation, we report the macro-F1 score to assess performance. For fairness analysis, we adopt Statistical Parity (SP), Equal Opportunity (EO), and Overall Accuracy Equality (OAE), as defined in Eqs. 1, 2, and 3, respectively.

Experimental Detail. For the evaluation of fairness, we adopt the `Qwen2-Audio-7B-Instruct` model[1]. To analyze the impact of the temperature parameter τ, we test the values $\{0.0, 0.3, 0.7, 1.0, 1.2, 1.5\}$. Dataset implementation follows the fold splits provided by EmoBox [29]. For each experimental configuration, we conduct 10 independent runs and report the mean and standard deviation of the results.

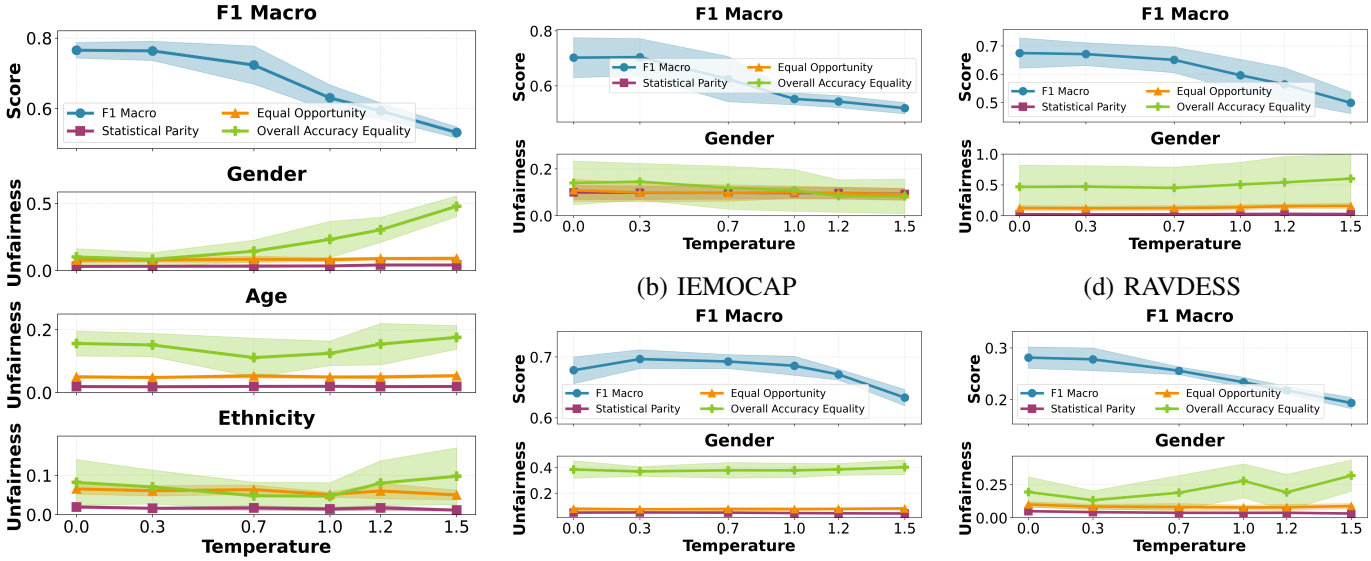
A. Zero-Shot SER Performance of ALMs

Table II reports the F1 performance of four ALMs—Qwen2-Audio, Audio-Flamingo-3 [9], SALMONN-7B [10], and Voxtral-Mini [11]—across five SER datasets, together with the official EmoBox baselines [29] obtained via fine-tuning.

ALM Comparison. Among the evaluated ALMs, Audio-Flamingo-3 remains below Qwen2-Audio on most datasets, with a stronger result only on EmoV-DB (83.85), while Voxtral-Mini and SALMONN-7B perform substantially worse (often below 32 F1). This variability highlights uneven zero-shot SER capability across ALMs; we therefore focus on Qwen2-Audio in the remainder of the paper.

Qwen2-Audio vs. Baselines. Operating fully zero-shot, Qwen2-Audio approaches the fine-tuned baselines on CREMA-D (76.56 vs. 76.60 for Whisper) and IEMOCAP (70.42 vs. 73.11), confirming effective zero-shot capture of emotion-relevant acoustic cues. Absolute scores are lower on EmoV-DB and MELD; on MELD, however, this reflects

[1]https://huggingface.co/Qwen/Qwen2-Audio-7B-Instruct



(a) CREMA-D (b) IEMOCAP (c) EmoV-DB (d) RAVDESS (e) MELD

Fig. 2: Performance and fairness of Qwen2-Audio at different temperatures. Top row: macro F1; remaining rows: fairness metrics—SP, EO, OAE—across sensitive attributes (Gender, Age, Ethnicity). Results averaged over folds; shaded areas show standard deviation.

TABLE IV: Fairness metrics—Statistical Parity (SP), Equal Opportunity (EO), and Overall Accuracy Equality (OAE)—across datasets and sensitive attributes.

Attribute	Dataset	SP↓	EO↓	OAE↓
Gender	CREMA-D	3.12	7.80	10.18
	IEMOCAP	9.84	9.92	14.54
	EmoV-DB	4.93	7.39	36.96
	RAVDESS	2.30	12.39	46.58
	MELD	4.89	9.93	19.27
Age	CREMA-D	1.92	5.01	15.62
Ethnicity	CREMA-D	1.95	6.54	8.19

the dataset's intrinsic difficulty rather than a model-specific weakness, as even the fine-tuned baselines remain below 33 F1 and Qwen2-Audio's 28.11 falls within their 24.99–32.95 range. The EmoBox baselines are fine-tuned with task-specific heads [29] and serve as a reference, not a direct comparison.

Effect of Postprocessing. Table III shows that the Levenshtein-based post-processing introduced in Section IV consistently improves F1 performance across datasets, confirming its effectiveness in normalizing generative ALM outputs for SER.

B. Fairness Results

Table IV reports the fairness metrics of Qwen2-Audio across gender, age, and ethnicity. All metrics are reported as percentages; lower is more equitable.

Across datasets, SP remains low for all sensitive attributes (gender SP ranges from 2.30 in RAVDESS to 9.84 in IEMOCAP), indicating that Qwen2-Audio produces similar overall prediction distributions across groups. EO is slightly higher, indicating moderate differences in error rates: gender disparities are larger in RAVDESS (12.39) and IEMOCAP (9.92), while age (5.01) and ethnicity (6.54) on CREMA-D are lower.

The largest disparities appear in OAE, indicating differences in classwise accuracy. Gender OAE is highest in RAVDESS (46.58) and EmoV-DB (36.96), with MELD at 19.27. For age and ethnicity, OAE is 15.62 and 8.19 (CREMA-D), respectively, showing that performance varies markedly across datasets and sensitive attributes. Concretely, SP, EO, and OAE are absolute between-group differences (in %) in positive-prediction rate, true-positive rate, and overall accuracy; a gender OAE of 46.6 on RAVDESS thus means accuracy differs by $\sim$47 points between groups, enough to cause systematically unequal reliability in deployment (e.g. emotion-aware healthcare). As metrics are computed per group, small groups and class imbalance inflate variance (large std), so values should be read within, not across, datasets.

C. Effect of Temperature on Performance and Fairness

Figure 2 illustrates how Qwen2-Audio's macro-F1 and fairness metrics evolve as the decoding temperature changes. Across all datasets, increasing the temperature degrades F1 Macro, with the most pronounced drops occurring beyond $\tau \geq 0.7$. This reflects the increased randomness in token generation, which harms the model's ability to produce consistent emotion labels. The fairness metrics respond unevenly: SP and EO remain largely stable, as added stochasticity perturbs overall output proportions and positive-class recall similarly across groups, whereas OAE fluctuates (notably on CREMA-D, RAVDESS, and MELD) because randomness interacts with each group's class-confusion structure, amplifying per-group

accuracy gaps. Temperature is thus a free, inference-time control over the accuracy–fairness trade-off, with OAE the metric most sensitive to it.

Overall, low-to-moderate temperatures ($\tau \leq 0.7$) achieve the best balance between accuracy and fairness, whereas higher values degrade both, reinforcing that controlled generation is crucial for fair zero-shot emotion recognition.

VI. Conclusion

We conducted a group-based fairness evaluation of Qwen2-Audio for zero-shot Speech Emotion Recognition using multigroup and multiclass fairness metrics across gender, age, and ethnicity. While the model achieves strong zero-shot performance and balanced output distributions, our analysis reveals notable disparities in error rates and classwise accuracy, particularly in terms of Overall Accuracy Equality, across datasets and sensitive attributes. We also find that disparities vary with generation temperature, making it a key hyperparameter for mitigating unfairness. In practice, this suggests auditing per-group accuracy (OAE), not only output parity (SP) which can mask disparities, and treating decoding temperature ($\tau \leq 0.7$) as a fairness-relevant hyperparameter; mitigation strategies for ALM-based SER remain an important future direction. A limitation of this work is the focus on English-only datasets; future work will extend the analysis to additional languages and ALMs.

Acknowledgment

This work has received funding from the European Union Horizon Research and Innovation programme under grant agreement No 101132389, REEVALUATE project, https://reevaluate.eu/.

References

[1] H. Lian, C. Lu, S. Li, Y. Zhao, C. Tang, and Y. Zong, "A survey of deep learning-based multimodal emotion recognition: Speech, text, and face," *Entropy*, vol. 25, no. 10, p. 1440, 2023.

[2] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[3] W. Kang, J. Jia, C. Wu, W. Zhou, E. Lakomkin, Y. Gaur, L. Sari, S. Kim, K. Li, J. Mahadeokar *et al.*, "Frozen large language models can perceive paralinguistic aspects of speech," *arXiv preprint arXiv:2410.01162*, 2024.

[4] Y. Li, Y. Gong, C.-H. H. Yang, P. Bell, and C. Lai, "Revise, reason, and recognize: Llm-based emotion recognition via emotion-specific prompts and asr error correction," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[5] Z. Cheng, Z.-Q. Cheng, J.-Y. He, K. Wang, Y. Lin, Z. Lian, X. Peng, and A. Hauptmann, "Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning," *Advances in Neural Information Processing Systems*, vol. 37, pp. 110 805–110 853, 2024.

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM computing surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.

[7] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic decision making and the cost of fairness," in *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, 2017, pp. 797–806.

[8] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[9] A. Goel, S. Ghosh, J. Kim, S. Kumar, Z. Kong, S.-g. Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle *et al.*, "Audio flamingo 3: Advancing audio intelligence with fully open large audio language models," *arXiv preprint arXiv:2507.08128*, 2025.

[10] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," *arXiv preprint arXiv:2310.13289*, 2023.

[11] A. H. Liu, A. Ehrenberg, A. Lo, C. Denoix, C. Barreau, G. Lample, J.-M. Delignon, K. R. Chandu, P. von Platen, P. R. Muddireddy *et al.*, "Voxtral," *arXiv preprint arXiv:2507.13264*, 2025.

[12] E. Purificato, L. Boratto, and E. W. De Luca, "Toward a responsible fairness analysis: from binary to multiclass and multigroup assessment in graph neural network-based user modeling tasks," *Minds and Machines*, vol. 34, no. 3, p. 33, 2024.

[13] T. M. Wani, T. S. Gunawan, S. A. A. Qadri, M. Kartiwi, and E. Ambikairajah, "A comprehensive review of speech emotion recognition systems," *IEEE access*, vol. 9, pp. 47 795–47 814, 2021.

[14] S. Akinpelu, S. Viriri, and A. Adegun, "An enhanced speech emotion recognition using vision transformer," *Scientific Reports*, vol. 14, no. 1, p. 13126, 2024.

[15] S. Liu, A. Mallol-Ragolta, E. Parada-Cabaleiro, K. Qian, X. Jing, A. Kathan, B. Hu, and B. W. Schuller, "Audio self-supervised learning: A survey," *Patterns*, vol. 3, no. 12, 2022.

[16] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[17] T. Feng, R. Hebbar, and S. Narayanan, "Trust-ser: On the trustworthiness of fine-tuning pre-trained speech embeddings for speech emotion recognition," in *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2024, pp. 11 201–11 205.

[18] M. M. Amin and B. W. Schuller, "On prompt sensitivity of chatgpt in affective computing," in *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*. IEEE, 2024, pp. 203–209.

[19] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[20] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2021.

[21] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," *Advances in neural information processing systems*, vol. 30, 2017.

[22] D. Pessach and E. Shmueli, "A review on fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 55, no. 3, pp. 1–44, 2022.

[23] I. Christop and M. Czajka, "Cameo: Collection of multilingual emotional speech corpora," *arXiv preprint arXiv:2505.11051*, 2025.

[24] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[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, no. 4, pp. 335–359, 2008.

[26] A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," *arXiv preprint arXiv:1806.09514*, 2018.

[27] S. R. Livingstone and F. A. Russo, "The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english," *PloS one*, vol. 13, no. 5, p. e0196391, 2018.

[28] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "Meld: A multimodal multi-party dataset for emotion recognition in conversations," in *Proceedings of the 57th annual meeting of the association for computational linguistics*, 2019, pp. 527–536.

[29] Z. Ma, M. Chen, H. Zhang, Z. Zheng, W. Chen, X. Li, J. Ye, X. Chen, and T. Hain, "Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark," *arXiv preprint arXiv:2406.07162*, 2024.