Multiclass Fairness Analysis of Qwen2-Audio in Speech Emotion Recognition

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 Qwen2Audio, a competitive ALM 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 Qwen2Audio 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 [19]. Recently, Audio-Language Models (ALMs) [11] 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 [16]. Others evaluate how prompts incorporating emotion-specific knowledge from acoustics, linguistics, and psychology can benefit the task [18]. Additional work proposes emotion-specific ALMs specialized through instruction tuning [9], or enhanced with contextual perception and chain-of-thought reasoning [32].

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 [24]. A common approach assesses fairness by exploiting sensitive attributes such as gender in a binary classification setting [12], [15]. 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 [11], [14], [20], [29] on SER and find that Qwen2-Audio achieves the best performance. Motivated by recent efforts to extend fairness assessment to multigroup–multiclass scenarios [27], we perform such an evaluation for SER using Qwen2-Audio on benchmarks including CREMA-D, IEMOCAP, EmoV-DB, RAVDESS, and MELD (see Table I).

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 [27] 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 [19], combined with traditional machine learning classifiers such as KNN, SVM, and Naïve Bayes [30]. The field progressed with the adoption of deep learning, from Convolutional Neural Networks [31] to Transformer-based models [2]. The emergence of Large Speech Models, pretrained on massive audio corpora using self-supervised objectives [21] or weak audio–text alignment signals [28], further advanced SER by providing powerful universal audio representations; many works build upon these models by training small classifiers atop frozen LSM features [13].

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 [11], [14], [20], [29], we select Qwen2-Audio as the best-performing model. Being a generative model, we evaluate it across various temperature settings [3] 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 [24]. 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 [4]. 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 [5], [12], [15], and *Individual Fairness*, which instead considers whether similar individuals receive similar model outputs [17].

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 [27] 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. According to recent reviews on machine learning fairness, *Statistical Parity*, *Equal Opportunity*, and *Overall Accuracy Equality* are well-defined and widely used fairness metrics [8], [25].

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 [12] 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 [15] requires equality of True Positive Rates between groups:

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

Overall Accuracy Equality (OAE). Overall Accuracy Equality [5] 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 take values in $[0, 1]$: they equal 0 under perfect fairness and increase as unfairness grows.

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	7 speakers	Gender	5
RAVDESS	24 (12M, 12F)	Gender	8
MELD	6 (3M, 3F)	Gender	7

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 [27], 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.

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 [28]) 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 $h'_A \in$

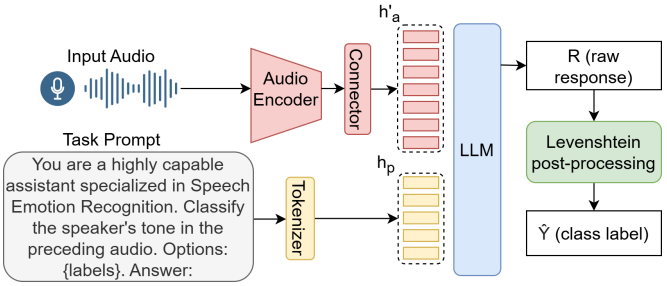


Fig. 1: Our Emotion Recognition pipeline exploiting ALMs.

$\mathbb{R}^{n'_A \times d_{\text{LLM}}}$ 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 [3], 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 [10], 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 [10].

V. Experiments

Datasets. We evaluate our approach on five widely used SER datasets (see Table I): *CREMA-D* [7], which contains 7,442 recordings from 96 actors spanning diverse ethnic, racial, and age groups; *IEMOCAP* [6], comprising 12 hours of emotionally annotated speech from ten actors across four emotion categories; *EmoV-DB* [1], which includes 6,887 recordings from four speakers labeled with five emotions; *RAVDESS* [22], consisting of 7,356 audio samples produced by 24 professional actors; and *MELD* [26] contains roughly 13,000 recordings from 1,433 *Friends* dialogues covering seven emotions. We limited our analysis to the six main characters, which account for 2,156 of 2,610 test samples, excluding others due to missing actor information.

TABLE II: F1 performance of large models on CREMA-D, IEMOCAP, EmoV-DB, RAVDESS, and TESS/MELD.

Model	CREMA-D	IEMOCAP	EmoV-DB	RAVDESS	MELD
HuBERT large	73.73	67.24	99.37	69.54	99.86
WavLM large	74.39	69.29	99.45	71.42	99.78
data2vec large	63.48	51.71	94.58	58.74	96.89
data2vec 2.0 large	69.25	56.70	98.19	70.94	99.54
Whisper large v3	76.60	73.11	99.34	75.19	99.96
Qwen2-Audio	76.56 (± 2.21)	70.42 (± 6.75)	69.64 (± 1.54)	67.48 (± 5.34)	28.11 (± 2.07)
Audio-Flamingo-3	61.00 (± 2.80)	46.95 (± 2.38)	83.85 (± 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.	76.56 (± 2.21)	70.42 (± 6.75)	69.64 (± 1.54)	28.11 (± 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 [23]. For each experimental configuration, we conduct 10 independent runs and report the mean and standard deviation of the results.

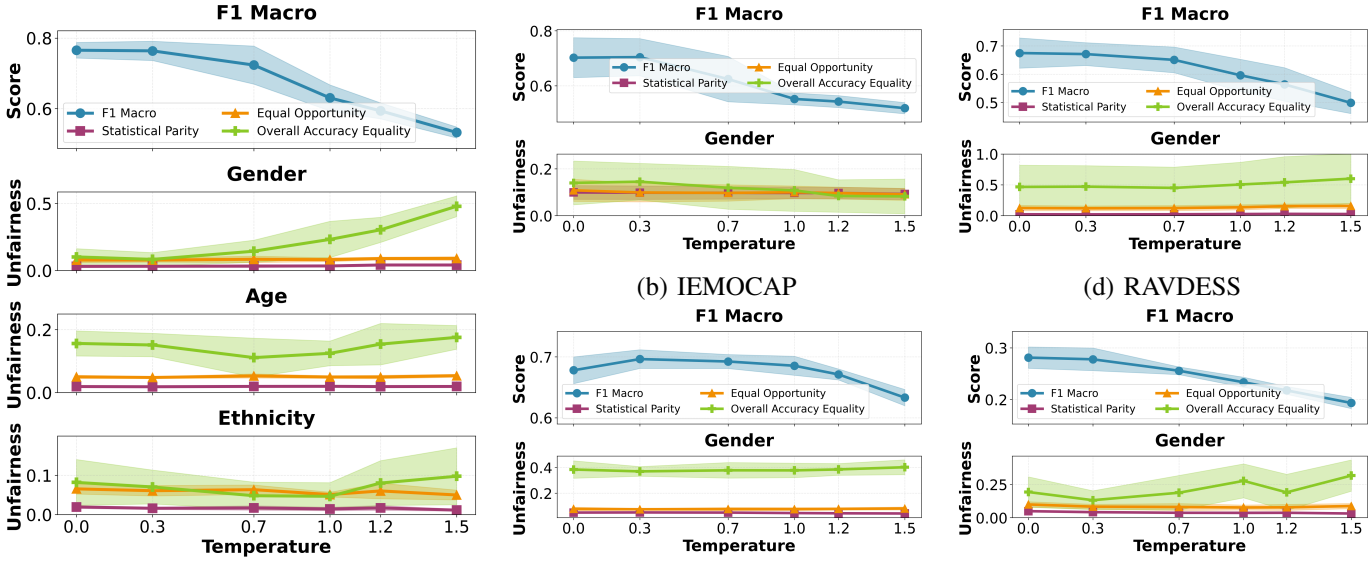
A. Zero-Shot SER Performance of ALMs

Table II reports the F1 performance of four ALMs—Qwen2-Audio, Audio-Flamingo-3 [14], SALMONN-7B [29], and Voxtral-Mini [20]—across five SER datasets, together with the official EmoBox baselines [23] obtained via fine-tuning.

ALM Comparison. Among the evaluated ALMs, Audio-Flamingo-3 shows moderate performance but remains below Qwen2-Audio on most datasets (e.g., 61.00 vs. 76.56 on CREMA-D and 46.95 vs. 70.42 on IEMOCAP), with a stronger result only on EmoV-DB (83.85). Voxtral-Mini and SALMONN-7B perform substantially worse overall, often scoring below 32 F1 (e.g., under 8 on CREMA-D and under 5 on RAVDESS). This variability highlights uneven zero-shot SER capability across ALMs; therefore, we focus on Qwen2-Audio in the remainder of the paper.

Qwen2-Audio vs. Baselines. Qwen2-Audio shows competitive results on CREMA-D, IEMOCAP, and RAVDESS, confirming effective zero-shot capture of emotion-relevant acoustic cues. Performance degrades on EmoV-DB due to higher speaker variability and subtler emotional expressions, and on MELD, whose conversational and multi-speaker nature

[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 Qwen2Audio 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

is less aligned with the model's general-purpose audio pretraining. The larger standard deviations on IEMOCAP (±6.75) and RAVDESS (±5.34) are mainly due to the higher number of cross-validation folds, which increase split-dependent variability. Overall, the results suggest that Qwen2-Audio encodes emotional information but remains less robust than speech-specialized models on certain SER benchmarks.

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 range from 0 to 100, with lower values indicating more equitable behavior.

Across datasets, SP remains relatively low for all sensitive attributes, suggesting that Qwen2-Audio produces similar overall prediction distributions across demographic groups. For example, SP ranges from 2.30 in RAVDESS to 9.84 in IEMOCAP for gender, indicating minimal overall bias in predictions. EO values are slightly higher, indicating moderate differences in error rates between groups. Gender disparities are more noticeable in RAVDESS (EO = 12.39) and IEMOCAP (EO = 9.92), while age shows a lower EO of 5.01 (CREMA-D) and ethnicity 6.54 (CREMA-D).

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 across datasets and sensitive attributes. This underscores that zero-shot emotion recognition with Qwen2-Audio remains uneven across sensitive attributes.

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 behavior reflects the increased randomness introduced in token generation, which harms the model's ability to produce consistent emotion labels. Fairness metrics exhibit different trends. SP and EO remain largely stable across all temperatures, suggesting that sampling stochasticity does not substantially affect output proportions across demographic groups. In contrast, OAE shows moderate fluctuations, particularly on CREMA-D, RAVDESS and MELD, indicating less consistent performance across groups.

Overall, low-to-moderate temperatures ($\tau \leq 0.7$) achieve the best balance between accuracy and fairness. As τ increases further, the model becomes less reliable in both prediction quality and equity, reinforcing that controlled generation is crucial for maintaining fair behavior in zero-shot emotion recognition.

VI. Conclusion

We conducted a group-based fairness evaluation of Qwen2-Audio for zero-shot Speech Emotion Recognition using multi-group and multiclass fairness metrics across gender, age, and ethnicity. While the model achieves competitive 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. A limitation of this work is the focus on English-only datasets; future work will extend the analysis to additional languages and ALMs.

References

[1] Adaeze Adigwe, Noé Tits, Kevin El Haddad, Sarah Ostadabbas, and Thierry Dutoit. The emotional voices database: Towards controlling the emotion dimension in voice generation systems. *arXiv preprint arXiv:1806.09514*, 2018.

[2] Samson Akinpelu, Serestina Viriri, and Adekanmi Adegun. An enhanced speech emotion recognition using vision transformer. *Scientific Reports*, 14(1):13126, 2024.

[3] Mostafa M Amin and Björn W Schuller. On prompt sensitivity of chatgpt in affective computing. In *2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 203–209. IEEE, 2024.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

[6] 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(4):335–359, 2008.

[7] Houwei Cao, David G Cooper, Michael K Keutmann, Ruben C Gur, Ani Nenkova, and Ragini Verma. Crema-d: Crowd-sourced emotional multimodal actors dataset. *IEEE transactions on affective computing*, 5(4):377–390, 2014.

[8] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *ACM Comput. Surv.*, 56(7), April 2024.

[9] Zebang Cheng, Zhi-Qi Cheng, Jun-Yan He, Kai Wang, Yuxiang Lin, Zheng Lian, Xiaojiang Peng, and Alexander Hauptmann. Emotion-llama: Multimodal emotion recognition and reasoning with instruction tuning. *Advances in Neural Information Processing Systems*, 37:110805–110853, 2024.

[10] Iwona Christop and Maciej Czajka. Cameo: Collection of multilingual emotional speech corpora. *arXiv preprint arXiv:2505.11051*, 2025.

[11] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017.

[13] Tiantian Feng, Rajat Hebbar, and Shrikanth 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)*, pages 11201–11205. IEEE, 2024.

[14] Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025.

[15] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[16] Wonjune Kang, Junteng Jia, Chunyang Wu, Wei Zhou, Egor Lakomkin, Yashesh Gaur, Leda Sari, Suyoun Kim, Ke Li, Jay Mahadeokar, et al. Frozen large language models can perceive paralinguistic aspects of speech. *arXiv preprint arXiv:2410.01162*, 2024.

[17] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *Advances in neural information processing systems*, 30, 2017.

[18] Yuanchao Li, Yuan Gong, Chao-Han Huck Yang, Peter Bell, and Catherine 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)*, pages 1–5. IEEE, 2025.

[19] Hailun Lian, Cheng Lu, Sunan Li, Yan Zhao, Chuangao Tang, and Yuan Zong. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face. *Entropy*, 25(10):1440, 2023.

[20] Alexander H Liu, Andy Ehrenberg, Andy Lo, Clément Denoix, Corentin Barreau, Guillaume Lample, Jean-Malo Delignon, Khyathi Raghavi Chandu, Patrick von Platen, Pavankumar Reddy Muddireddy, et al. Voxtral. *arXiv preprint arXiv:2507.13264*, 2025.

[21] Shuo Liu, Adria Mallol-Ragolta, Emilia Parada-Cabaleiro, Kun Qian, Xin Jing, Alexander Kathan, Bin Hu, and Bjoern W Schuller. Audio self-supervised learning: A survey. *Patterns*, 3(12), 2022.

[22] Steven R Livingstone and Frank 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*, 13(5):e0196391, 2018.

[23] Ziyang Ma, Mingjie Chen, Hezhao Zhang, Zhisheng Zheng, Wenxi Chen, Xiquan Li, Jiaxin Ye, Xie Chen, and Thomas Hain. Emobox: Multilingual multi-corpus speech emotion recognition toolkit and benchmark. *arXiv preprint arXiv:2406.07162*, 2024.

[24] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[25] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[26] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada 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*, pages 527–536, 2019.

[27] Erasmo Purificato, Ludovico Boratto, and Ernesto William 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*, 34(3):33, 2024.

[28] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR, 2023.

[29] Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*, 2023.

[30] Taiba Majid Wani, Teddy Surya Gunawan, Syed Asif Ahmad Qadri, Mira Kartiwi, and Eliathamby Ambikairajah. A comprehensive review of speech emotion recognition systems. *IEEE access*, 9:47795–47814, 2021.

[31] Jianfeng Zhao, Xia Mao, and Lijiang Chen. Speech emotion recognition using deep 1d & 2d cnn lstm networks. *Biomedical signal processing and control*, 47:312–323, 2019.

[32] Zhixian Zhao, Xinfa Zhu, Xinsheng Wang, Shuiyuan Wang, Xuelong Geng, Wenjie Tian, and Lei Xie. Steering language model to stable speech emotion recognition via contextual perception and chain of thought. *arXiv preprint arXiv:2502.18186*, 2025.