ERM-MinMaxGAP: Benchmarking and Mitigating Gender Bias in Multilingual Multimodal Speech-LLM Emotion Recognition

Zi Haur Pang [1] **Xiaoxue Gao** [2] **Tatsuya Kawahara** [1] **Nancy F. Chen** [2]

Abstract

Speech emotion recognition (SER) systems exhibit gender-related performance disparities, but how such bias manifests in multilingual speech LLMs across languages and modalities is unclear. We introduce a novel multilingual, multimodal benchmark built on MELD-ST, spanning English, Japanese, and German, to quantify language-specific SER performance and gender gaps. We find bias is strongly language-dependent, and multimodal fusion does not reliably improve fairness. To address these, we propose ERM-MinMaxGAP, a fairness-informed training objective, which augments empirical risk minimization (ERM) with a proposed adaptive fairness weight mechanism and a novel MinMaxGAP regularizer on the maximum male-female loss gap within each language and modality. Building upon the Qwen2-Audio backbone, our ERM-MinMaxGAP improves multilingual SER performance by 5.5% and 5.0% while reducing the overall gender bias gap by 0.1% and 1.4% in the unimodal and multimodal settings, respectively.

1. Introduction

Speech emotion recognition (SER) (Schuller, 2018) supports emotion-aware agents (Sanjeewa et al., 2024), affective tutoring (Petrovica et al., 2017), call-center analytics (Martín-Doñas et al., 2024), and mental-health assessment (Jordan et al., 2025; Pang et al., 2026), and is commonly evaluated on corpora such as IEMOCAP (Busso et al., 2008) and MSP-IMPROV (Busso et al., 2016). Recent self-supervised speech models have improved SER performance (Gao et al., 2023; Pepino et al., 2021), while end-to-end audio–language and speech-large language models (LLMs), are emerging as general-purpose speech reasoning systems (Bellver Soler et al., 2024; Chu et al., 2024; Zhang et al., 2023). However, robust SER remains difficult due to speaker variability (Mariooryad & Busso, 2014; Sethu et al., 2013), annotation subjectivity (Tavernor et al., 2024), domain mismatch (Pastor et al., 2024), and fairness concerns: models may exploit demographic or linguistic shortcuts in acoustic–prosodic patterns, yielding uneven performance across speaker groups (Mariooryad & Busso, 2014; Ulgen et al., 2024). Similar disparities have been observed in ASR across race, gender, dialect, and language (Attanasio et al., 2024; Harris et al., 2024; Koenecke et al., 2020; Tatman, 2017; Veliche et al., 2024), and SER studies have reported gender gaps and fairness–accuracy trade-offs (Gorrostieta et al., 2019; Lin et al., 2025).

Despite this progress, SER fairness research has mainly studied classifier-style systems built on fixed SSL representations (Lin et al., 2025), while fairness analyses of speech-integrated LLMs have focused mostly on semantic tasks rather than emotion recognition (Lin et al., 2024a). Thus, gender bias in multilingual, multimodal SER with speech LLMs remains under-benchmarked. To address this gap, we present a controlled benchmark that disentangles language from corpus effects, and propose **ERM-MinMaxGAP**, a fairness-aware objective that augments empirical risk minimization with a penalty on the maximum within-language, within-modality male–female loss gap. Our contributions are: (1) the first dedicated benchmark of gender bias in multilingual, multimodal speech LLMs for SER; and (2) ERM-MinMaxGAP, which optimizes SER loss while directly reducing the worst gender disparity. Project page: `https://github.com/zihaurpang/ERM-MinMaxGAP`.

2. Methodology

2.1. ERM-Based Supervised Fine-Tuning

As our main objective is to optimize SER, we start with an empirical risk minimization (ERM) procedure, i.e., minimizing the average loss over the training set. In practice, we apply supervised fine-tuning (SFT) with Low-Rank Adaptation (LoRA) adaptation (Hu et al., 2022) in this study.

[1]Kyoto University, Japan [2]Agency for Science, Technology, and Research (A*STAR), Singapore. Correspondence to: Xiaoxue Gao <{Gao_Xiaoxue}@a-star.edu.sg>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026.

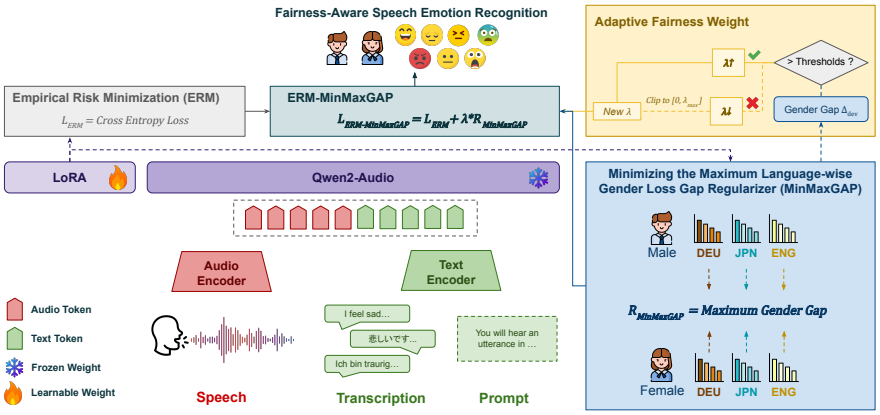

Figure 1. Architecture of the proposed method. The method consists of (1) empirical risk minimization for overall SER improvement, (2) MinMaxGAP for minimizing the language-wise gender gap, and (3) adaptive fairness-weight adjustment for fairness-aware SER.

We use the cross-entropy loss (Zhang & Sabuncu, 2018) as our main objective function, denoted as $\mathcal{L}_{\mathrm{ERM}}(\theta) = CrossEntropyLoss$, where θ represents the model parameters, as shown in the top-left part of Figure 1.

2.2. MinMaxGAP Regularization

To reduce gender disparity in multilingual SER, we introduce **Minimizing the Maximum Language-wise Gender Loss Gap Regularizer (MinMaxGAP)**, a fairness regularizer that explicitly measures the male–female loss gap *within each language*, as illustrated in the bottom-right part of Figure 1. To measure the maximum gender gap within each language, we first define the conditional mean loss $\mathcal{L}_{\ell,g}(\theta) = \mathbb{E}[\mathcal{L}_i(\theta) \mid \ell_i = \ell,\ g_i = g]$, where $\ell \in \mathcal{L}$ denotes the language and $g \in \{\mathrm{F}, \mathrm{M}\}$ denotes the gender group. The within-language gender gap for language ℓ is then defined as $\Delta_\ell(\theta) = |\mathcal{L}_{\ell,\mathrm{F}}(\theta) - \mathcal{L}_{\ell,\mathrm{M}}(\theta)|$. To prevent a large disparity in one language from being masked by smaller disparities in others, instead of averaging the gaps across languages, we focus on the worst-case language: $\Delta_{\max}(\theta) = \max_{\ell \in \mathcal{L}} \Delta_\ell(\theta)$. We further include a fairness regularizer that penalizes the training objective, defined as $\mathcal{R}_{\mathrm{MinMaxGAP}}(\theta) = (\Delta_{\max}(\theta))^p$, where $p \in \{1, 2\}$ controls the penalty shape. In our main setting, we use $p = 2$, which places a stronger penalty on large disparities.

2.3. Adaptive Fairness Weight

A fixed fairness weight is often suboptimal across training stages, especially early in training: the model should prioritize learning the fundamental SER task, whereas stronger fairness pressure may become more beneficial once the model reaches a reasonable level of performance. To address this, we propose a dynamic fairness adjustment mechanism that adaptively updates the fairness weight based on the development-set gap. Drawing inspiration from constrained optimization (Bertsekas, 2014), we employ a Lagrange multiplier method to refine the weight adaptively during training (Figure 1 top right). We update the fairness weight as $\lambda^{(k+1)} = \Pi_{[0,\lambda_{\max}]}\left(\lambda^{(k)} + \eta\left(\Delta_{\mathrm{dev}}^{(k)} - \epsilon\right)\right)$, where $\Delta_{\mathrm{dev}}^{(k)}$ denotes the fairness gap measured on the development set at evaluation step k, ϵ is the target tolerance, η is the update rate, $\lambda_{\max}$ is the maximum regularization strength, and $\Pi_{[0,\lambda_{\max}]}$ denotes clipping onto the interval $[0, \lambda_{\max}]$. Intuitively, when the observed development-set gap exceeds the target threshold ϵ, the regularization weight increases, placing more emphasis on reducing disparity. When the gap below the threshold, the update reduces or stops, allowing the optimization to focus more on task performance.

2.4. Final Training Objective

As illustrated in the top-middle part of Figure 1, we combine the ERM term and the fairness regularizer to form the final training objective. At step k, the objective is defined as $\mathcal{L}_{\mathrm{ERM\text{-}MinMaxGAP}}^{(k)}(\theta) = \mathcal{L}_{\mathrm{ERM}}(\theta) + \lambda^{(k)}\,\mathcal{R}_{\mathrm{MinMaxGAP}}(\theta)$. This objective jointly optimizes overall SER performance and worst-language gender fairness. The ERM term preserves general recognition ability, while the MinMaxGAP term directly suppresses the largest male–female loss disparity across languages.

3. Experimental Setup

Database We conduct experiments on **MELD-ST** (Chen et al., 2024), a multilingual emotion-aware speech dataset derived from MELD (Poria et al., 2019). It extends the original English dialogue data with aligned Japanese and German speech pairs while preserving 7-class emotion labels. To support gender-fairness analysis, we additionally annotate speaker gender, with statistics shown in Table 1.

Training Hyperparameters We fine-tune **Qwen2-Audio-7B-Instruct** (Chu et al., 2024) on MELD-ST for up to 20

Table 1. Statistics of the MELD-ST dataset.

Language	Gender	Train	Valid	Test	Total
English	*Female*	3546	464	501	4511
	Male	3850	448	459	4757
	Total	7396	912	960	9268
Japanese	*Female*	3546	464	501	4511
	Male	3850	448	459	4757
	Total	7396	912	960	9268
German	*Female*	4151	546	525	5222
	Male	4425	514	555	5494
	Total	8576	1060	1080	10716
Overall	*Female*	11243	1474	1527	14244
	Male	12125	1410	1473	15008
	Total	23368	2884	3000	29252

epochs using an effective batch size of 64, learning rate 5×10^{-5}, weight decay 0.01, early stopping patience 5, and seed 42. For LoRA (Hu et al., 2022), we use $r = 16$, $\alpha = 32$, and dropout 0.05. For the fairness objective, we set $p = 2$, initialize $\lambda = 0$, and adaptively update it with $\epsilon = 0.02$, update rate 0.5, and maximum $\lambda = 10.0$.

Comparison Models We evaluate **unimodal**, where the model receives speech and the task instruction only, and **multimodal**, where it additionally receives the ground-truth transcription. We compare our method with recent speech LLMs in the zero-shot setting: Qwen2-Audio-7B-Instruct (Chu et al., 2024), Voxtral-Mini-3B (Liu et al., 2025), gpt4o-mini-audio[1], Kimi-Audio-7B-Instruct (Team, 2024), and Ultravox-0.4[2].

Evaluation Metrics We evaluate both **SER performance** and **gender fairness** (Lin et al., 2024b; 2025). For SER, we report **weighted F1 (W-F1)** and **accuracy (ACC)**, where higher is better. For fairness, we report gender gaps in **True Positive Rate (TPR)**, **False Positive Rate (FPR)**, **W-F1**, and **ACC**. TPR/FPR gaps are computed in a one-vs-rest manner and averaged over emotion classes, while W-F1/ACC gaps measure absolute performance differences between female and male speakers. Lower values indicate smaller disparity. We further report **AVG**, the mean of the gap metrics, as a compact summary of overall gender bias.

4. Results and Analysis

Benchmarking Gender Bias in Multilingual Multimodal Speech LLMs Table 2 benchmarks gender bias together with overall SER performance for recent speech LLMs across languages and modalities. We find that multimodal input often improves SER, but does not consistently improve fairness. Some models (e.g., kimi-audio-7b) show gains in both SER and reduced gender gap, whereas others improve overall SER while exhibiting a larger gender gap, indicating that multimodal fusion can improve recognition accuracy without reliably reducing gender disparity. The benchmark also shows strong language dependence. English is generally the easiest setting and Japanese the most difficult, with stronger models achieving higher W-F1/ACC in ENG than in JPN under both input conditions. Fairness trends are likewise inconsistent across languages: a model may reduce the gender gap in one language but enlarge it in another. Overall, gender bias in multilingual multimodal SER is highly model- and language-dependent, motivating a fairness-aware objective rather than relying on multimodal fusion alone.

Effectiveness of ERM-MinMaxGAP As shown in Table 2, **ERM-MinMaxGAP** achieves the best overall performance while consistently reducing gender disparity. In the multilingual setting, it obtains the best SER in both conditions, with gains of +5.49 W-F1 and +9.75 ACC in the unimodal setting, and +5.03 W-F1 and +3.62 ACC in the multimodal setting, relative to the best baseline, while maintaining the second-smallest gender gap. Importantly, its AVG gender gap is also reduced by 0.8 when moving to multimodal input.

In the monolingual settings, ERM-MinMaxGAP remains the strongest overall performer in all languages while keeping gender disparity as small as possible. It is also observed that it provides consistent mitigation of the AVG gender gap in each language. These results show that ERM-MinMaxGAP improves SER while delivering a stronger overall performance–fairness trade-off across languages and modalities. Although it does not always yield the minimum post-hoc gap in every setting, this is consistent with its goal of penalizing the worst within-language group disparity rather than directly minimizing an average fairness metric.

Ablation Study Table 3 confirms the contribution of each component to **ERM-MinMaxGAP**. Relative to zero-shot Qwen2-Audio, ERM-based supervised fine-tuning (SFT) already provides a large gain while reducing the AVG gender gap in both the unimodal and multimodal settings. Moreover, our proposed MinMaxGAP further improves overall SER performance while reducing gender disparity in both settings. This suggests that MinMaxGAP improves the overall performance–fairness trade-off rather than uniformly minimizing every post-hoc gap metric. We also evaluated a MinMaxGAP-only ablation, but without the main task objective (ERM), the model could not produce a valid single prediction, and thus being excluded from the study.

We also evaluated the effect of the fairness weight λ. Compared with our proposed adaptive weight update method, a fixed λ reduces the gender gap as it increases, but also degrades SER, revealing a clear fairness–utility trade-off. We also examined whether the penalty power affects overall

[1] https://developers.openai.com/api/docs/models/gpt-4o-mini-audio-preview

[2] https://github.com/fixie-ai/ultravox

Table 2. Comparison of unimodal and multimodal performance, with SER results and gender bias gap metrics [%]. Relative improvements and degradations of the multimodal input over the unimodal input are shown in green, and red, respectively. Top two results are highlighted in **bold** and underline, respectively.

Model	Unimodal (Speech only)							Multimodal (Speech + Transcription)						
	SER Result		Gender Bias Gap					SER Result		Gender Bias Gap				
	W-F1↑	ACC↑	TPR↓	FPR↓	W-F1↓	ACC↓	AVG↓	W-F1↑	ACC↑	TPR↓	FPR↓	W-F1↓	ACC↓	AVG↓
Multilingual														
Qwen2-Audio	34.89	33.31	9.78	4.26	3.36	4.67	5.51	34.62	30.79	8.66	2.08	3.67	3.35	4.44 ↓1.07
Voxtral-Mini-3B	44.78	44.52	9.28	2.74	4.97	5.22	5.55	50.04	55.03	7.27	1.90	2.18	**1.84**	**3.30** ↓2.25
gpt4o-mini-audio	45.89	44.57	9.61	**1.48**	2.61	4.03	4.43	52.65	51.76	8.09	2.08	4.38	5.25	4.95 ↑0.52
kimi-audio-7b	40.27	39.96	9.18	3.12	5.21	3.58	5.27	42.34	42.56	7.15	2.36	3.21	3.26	4.00 ↓1.27
Ultravox-0.4	27.43	25.29	**2.76**	1.49	**2.41**	**1.09**	**1.94**	32.45	30.78	7.79	**1.86**	5.65	4.42	4.93 ↑2.99
ERM-MinMaxGAP (Ours)	**51.38**	**54.32**	6.38	3.01	3.52	4.44	4.34	**57.68**	**58.65**	**7.08**	2.69	**1.84**	2.53	3.53 ↓0.80
Monolingual														
DEU														
Qwen2-Audio	31.02	28.15	12.05	5.72	4.50	3.25	6.38	32.18	26.85	10.25	1.96	3.22	1.49	4.23 ↓2.15
Voxtral-Mini-3B	45.71	45.65	9.60	2.50	4.09	4.21	5.10	48.46	53.33	7.14	1.47	0.95	**0.00**	2.39 ↓2.71
gpt4o-mini-audio	46.55	44.54	8.59	1.66	**1.60**	3.03	3.72	52.40	51.20	10.13	1.98	5.39	6.37	5.97 ↑2.25
kimi-audio-7b	39.03	38.43	10.75	3.04	4.27	2.13	5.05	44.16	42.69	**4.14**	1.67	4.53	5.15	3.87 ↓1.17
Ultravox-0.4	27.97	25.56	**1.68**	**0.50**	1.92	**0.43**	**1.13**	31.71	28.89	4.97	**0.67**	**0.46**	1.61	**1.93** ↑0.79
ERM-MinMaxGAP (Ours)	**47.84**	**51.39**	3.06	3.71	7.47	9.19	5.86	**53.32**	**54.91**	8.01	3.36	3.92	4.92	5.05 ↓0.81
ENG														
Qwen2-Audio	46.02	44.27	7.69	3.65	0.92	1.75	3.50	46.13	41.67	10.34	2.31	4.57	5.11	5.58 ↑2.08
Voxtral-Mini-3B	41.98	38.75	9.88	2.75	7.57	7.87	7.02	53.74	57.71	7.64	2.04	3.57	2.46	3.93 ↓3.09
gpt4o-mini-audio	49.67	49.06	12.47	**1.52**	4.22	5.93	6.03	56.99	56.35	5.77	**1.36**	4.42	5.29	4.21 ↓1.82
kimi-audio-7b	57.86	54.79	10.33	1.54	5.52	5.63	5.80	56.11	53.85	10.58	1.71	2.02	2.17	4.77 ↓0.80
Ultravox-0.4	35.80	32.50	**4.01**	2.43	1.76	**0.49**	**2.17**	39.98	38.44	9.56	1.88	8.93	7.28	6.91 ↑4.74
ERM-MinMaxGAP (Ours)	**58.09**	**60.52**	8.84	2.10	**0.41**	0.92	3.07	**68.13**	**68.85**	**5.08**	1.74	**0.38**	**1.24**	**2.11** ↓0.96
JPN														
Qwen2-Audio	27.64	27.50	9.07	2.88	3.56	7.19	5.68	25.56	23.85	**3.61**	**1.95**	3.03	2.29	**2.72** ↓2.96
Voxtral-Mini-3B	46.63	49.17	8.27	2.96	0.20	**1.39**	3.21	47.91	**54.06**	7.01	2.12	**0.81**	2.03	2.99 ↓0.22
gpt4o-mini-audio	41.45	40.10	6.89	1.24	**0.11**	2.12	2.59	48.56	47.71	7.77	2.69	2.99	3.75	4.30 ↑1.71
kimi-audio-7b	23.91	26.67	5.50	4.18	5.72	1.50	4.23	26.76	31.15	4.95	3.32	2.52	**0.82**	2.90 ↓1.32
Ultravox-0.4	18.52	17.81	**2.01**	**0.74**	3.25	1.77	**1.94**	25.66	25.00	8.12	2.52	3.99	1.77	4.10 ↑2.16
ERM-MinMaxGAP (Ours)	**48.19**	**51.04**	7.23	3.21	2.68	3.22	4.09	**51.58**	52.19	8.13	2.97	1.22	1.44	3.44 ↓0.64

Table 3. Ablation study [%]. Relative improvements and degradations of the multimodal input over the unimodal input are shown in green, and red, respectively.

Model	Unimodal (Speech only)							Multimodal (Speech + Transcription)						
	SER Result		Gender Bias Gap					SER Result		Gender Bias Gap				
	W-F1↑	ACC↑	TPR↓	FPR↓	W-F1↓	ACC↓	AVG↓	W-F1↑	ACC↑	TPR↓	FPR↓	W-F1↓	ACC↓	AVG↓
Main Components														
Qwen2-Audio	34.89	33.31	9.78	4.26	3.36	4.67	5.51	34.62	30.79	8.66	2.08	3.67	3.35	4.44 ↓1.07
+ ERM (SFT)	47.50	46.07	11.28	2.64	2.87	3.08	4.97	56.13	54.77	9.50	2.03	0.93	1.22	3.42 ↓1.55
ERM-MinMaxGAP (Ours)	51.38	54.32	6.38	3.01	3.52	4.44	4.34	57.68	58.65	7.08	2.69	1.84	2.53	3.53 ↓0.80
Lambda Effect														
$\lambda = 0$ (SFT)	47.50	46.07	11.28	2.64	2.87	3.08	4.97	56.13	54.77	9.50	2.03	0.93	1.22	3.42 ↓1.55
$\lambda = 1$	32.71	37.86	3.31	1.59	5.73	4.33	3.74	47.61	45.81	7.62	1.38	1.17	1.59	2.94 ↓0.80
$\lambda = 5$	30.97	30.17	3.83	0.72	3.33	3.15	2.76	35.50	33.31	6.04	1.31	2.03	2.12	2.87 ↑0.12
$\lambda = 10$	29.67	28.77	3.87	1.23	2.23	2.55	2.47	30.87	31.04	3.12	0.68	2.92	3.35	2.52 ↑0.05
$\lambda =$ adaptive (Ours)	51.38	54.32	6.38	3.01	3.52	4.44	4.34	57.68	58.65	7.08	2.69	1.84	2.53	3.53 ↓0.80
Penalty Power Effect														
$p = 1$	51.38	54.32	6.38	3.01	3.52	4.44	4.34	58.76	59.77	7.91	2.81	1.91	2.30	3.73 ↓0.61
$p = 2$ (Ours)	51.38	54.32	6.38	3.01	3.52	4.44	4.34	57.68	58.65	7.08	2.69	1.84	2.53	3.53 ↓0.80

performance. We find that although $p = 1$ yields slightly higher multimodal SER, it also produces a larger gender gap, whereas our proposed $p = 2$ provides better fairness.

5. Conclusion

This paper presented, to the best of our knowledge, the first dedicated benchmark of gender bias in multilingual multimodal speech LLMs for SER. Through evaluation on MELD-ST across English, Japanese, and German, we showed that gender disparity is highly dependent on both language and input modality, and that multimodal input does not reliably reduce bias even when it improves SER performance. To address this issue, we proposed **ERM-MinMaxGAP**, a fairness-aware training objective that combines empirical risk minimization with worst-language gender-gap regularization. Experimental results showed that ERM-MinMaxGAP achieves the strongest overall performance while providing more consistent gender-gap mitigation across languages and modalities, yielding a better performance–fairness trade-off than the compared baselines.

References

Attanasio, G., Savoldi, B., Fucci, D., and Hovy, D. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 21318–21340, 2024.

Bellver Soler, J., Martín Fernández, I., Bravo Pacheco, J. M., Esteban Romero, S., Fernández Martínez, F., and D'Haro Enríquez, L. F. Multimodal audio-language model for speech emotion recognition. *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024.

Bertsekas, D. P. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

Busso, C., Bulut, M., Lee, C.-C., Kazemzadeh, A., Mower, E., Kim, S., Chang, J. N., Lee, S., and Narayanan, S. S. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4): 335–359, 2008.

Busso, C., Parthasarathy, S., Burmania, A., AbdelWahab, M., Sadoughi, N., and Provost, E. M. Msp-improv: An acted corpus of dyadic interactions to study emotion perception. *IEEE transactions on affective computing*, 8(1): 67–80, 2016.

Chen, S., Yahata, S., Shimizu, S., Yang, Z., Li, Y., Chu, C., and Kurohashi, S. Meld-st: An emotion-aware speech translation dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 10118–10126, 2024.

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

Gao, Y., Chu, C., and Kawahara, T. Two-stage Finetuning of Wav2vec 2.0 for Speech Emotion Recognition with ASR and Gender Pretraining. In *Interspeech 2023*, pp. 3637–3641, 2023. doi: 10.21437/Interspeech.2023-756.

Gorrostieta, C., Lotfian, R., Taylor, K., Brutti, R., and Kane, J. Gender de-biasing in speech emotion recognition. In *Interspeech*, pp. 2823–2827, 2019.

Harris, C., Mgbahurike, C., Kumar, N., and Yang, D. Modeling gender and dialect bias in automatic speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15166–15184, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.

Jordan, E., Terrisse, R., Lucarini, V., Alrahabi, M., Krebs, M.-O., Desclés, J., and Lemey, C. Speech emotion recognition in mental health: Systematic review of voice-based applications. *JMIR mental health*, 12(1):e74260, 2025.

Koenecke, A., Nam, A., Lake, E., Nudell, J., Quartey, M., Mengesha, Z., Toups, C., Rickford, J. R., Jurafsky, D., and Goel, S. Racial disparities in automated speech recognition. *Proceedings of the national academy of sciences*, 117(14):7684–7689, 2020.

Lin, Y.-C., Lin, T.-Q., Yang, C.-K., Lu, K.-H., Chen, W.-C., Kuan, C.-Y., and Lee, H.-y. Listen and speak fairly: a study on semantic gender bias in speech integrated large language models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 439–446. IEEE, 2024a.

Lin, Y.-C., Wu, H., Chou, H.-C., Lee, C.-C., and Lee, H.-y. Emo-bias: A large scale evaluation of social bias on speech emotion recognition. *INTERSPEECH 2024*, 2024b.

Lin, Y.-C., Chou, H.-C., Liang, Y.-H. L., and Lee, H.-y. Emo-debias: Benchmarking gender debiasing techniques in multi-label speech emotion recognition. *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2025)*, 2025.

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

Mariooryad, S. and Busso, C. Compensating for speaker or lexical variabilities in speech for emotion recognition. *Speech Communication*, 57:1–12, 2014.

Martín-Doñas, J. M., Zorrilla, A. L., deVelasco, M., Vásquez-Correa, J. C., Álvarez, A., Torres, M. I., Delgado, P., Lazpiur, A., Romero, B., and Alkorta, I. Speech emotion recognition for call centers using self-supervised models: A complete pipeline for industrial applications. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pp. 119–128, 2024.

Pang, Z., Fu, Y., Gao, Y., and Kawahara, T. Paralinguistic emotion-aware validation timing detection in japanese empathetic spoken dialogue. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

Pastor, M. A., Ortega, A., and Ribas, D. Analysis of the domain mismatch problem in the speech emotion recognition task. In *Proc. IberSPEECH 2024*, pp. 181–185, 2024.

Pepino, L., Riera, P., and Ferrer, L. Emotion Recognition from Speech Using wav2vec 2.0 Embeddings. In *Interspeech 2021*, pp. 3400–3404, 2021. doi: 10.21437/Interspeech.2021-703.

Petrovica, S., Anohina-Naumeca, A., and Ekenel, H. K. Emotion recognition in affective tutoring systems: Collection of ground-truth data. *Procedia Computer Science*, 104:437–444, 2017.

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

Sanjeewa, R., Iyer, R., Apputhurai, P., Wickramasinghe, N., and Meyer, D. Empathic conversational agent platform designs and their evaluation in the context of mental health: systematic review. *JMIR Mental Health*, 11: e58974, 2024.

Schuller, B. W. Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends. *Communications of the ACM*, 61(5):90–99, 2018.

Sethu, V., Epps, J., and Ambikairajah, E. Speaker variability in speech based emotion models-analysis and normalisation. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 7522–7526. IEEE, 2013.

Tatman, R. Gender and dialect bias in YouTube's automatic captions. In Hovy, D., Spruit, S., Mitchell, M., Bender, E. M., Strube, M., and Wallach, H. (eds.), *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pp. 53–59, Valencia, Spain, April 2017. Association for Computational Linguistics.

Tavernor, J., El-Tawil, Y., and Provost, E. M. The whole is bigger than the sum of its parts: Modeling individual annotators to capture emotional variability. *Interspeech 2024*, 2024.

Team, K. Kimi-audio technical report, 2024.

Ulgen, I. R., Du, Z., Busso, C., and Sisman, B. Revealing emotional clusters in speaker embeddings: A contrastive learning strategy for speech emotion recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12081–12085. IEEE, 2024.

Veliche, I.-E., Huang, Z., Kochaniyan, V. A., Peng, F., Kalinli, O., and Seltzer, M. L. Towards measuring fairness in speech recognition: Fair-speech dataset. *Interspeech 2024*, 2024.

Zhang, D., Li, S., Zhang, X., Zhan, J., Wang, P., Zhou, Y., and Qiu, X. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In Bouamor, H., Pino, J., and Bali, K. (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 15757–15773, Singapore, December 2023. Association for Computational Linguistics.

Zhang, Z. and Sabuncu, M. Generalized cross entropy loss for training deep neural networks with noisy labels. *Advances in neural information processing systems*, 31, 2018.