

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

Zi Haur Pang¹, Xiaoxue Gao^{2,**}, Tatsuya Kawahara¹, Nancy F. Chen²

¹ Kyoto University, Japan

² Agency for Science, Technology, and Research (A*STAR), Singapore

{pang, kawahara}@sap.ist.i.kyoto-u.ac.jp, {Gao.Xiaoxue, Nancy.Chen}@a-star.edu.sg

Abstract

Speech emotion recognition (SER) systems can 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 approach 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.

Index Terms: Speech Emotion Recognition, Speech LLMs, Fairness, Bias, Multilingual LLMs, Multimodal LLMs

1. Introduction

Speech emotion recognition (SER) [1] is a key capability in affective computing and spoken human-computer interaction, enabling emotion-aware conversational agents [2], affective tutoring [3], call-center analytics [4], and mental-health assessment [5, 6]. SER is commonly benchmarked on corpora such as IEMOCAP [7] and MSP-IMPROV [8]. Recent advances in self-supervised speech pretraining have substantially improved SER by transferring representations from large pretrained encoders [9–12]. In parallel, the community is rapidly moving toward end-to-end audio-language and speech-large language models (LLMs) architectures that follow natural-language instructions and integrate audio with text, bringing SER closer to general-purpose multimodal interaction [13–17]. Despite these gains, robust deployment remains difficult due to speaker variability [18, 19], annotation subjectivity [20], and domain mismatch across corpora and recording conditions [21].

A particularly important challenge is fairness. Because emotions are expressed through acoustic-prosodic patterns that correlate with speaker traits and language, SER models may exploit demographic or linguistic shortcuts rather than emotion-relevant cues, leading to uneven performance across groups [19, 22]. Fairness issues have also been repeatedly studied in related speech technologies, including racial disparities in commercial automatic speech recognition (ASR) [23], gender and

dialect bias in English ASR [24, 25], and persistent gender performance gaps in multilingual speech recognition across many languages [26, 27]. In SER specifically, prior work reported measurable gender disparities and showed that debiasing can improve consistency between female and male speakers [28], while more recent benchmarking indicates that mitigating disparity often involves a nontrivial fairness-accuracy trade-off, and that debiasing behavior can depend strongly on dataset composition and method choice [29].

However, existing SER fairness studies mainly focus on classifier-style pipelines built on fixed self-supervised learning (SSL) representations rather than modern end-to-end audio-language or speech-LLM systems [29]. This gap matters because speech LLMs are increasingly used as general audio reasoning backbones [13–16]. Meanwhile, fairness analyses of speech-integrated LLMs have largely centered on semantic tasks such as speech-to-text translation, spoken coreference, sentence continuation, and spoken question answering, rather than SER [30]. As a result, gender bias in multilingual, multimodal SER with speech LLMs remains under-benchmarked.

To fill this gap, we present a dedicated benchmark of gender bias in multilingual, multimodal speech LLMs for SER under a controlled setting that disentangles language from corpus effects. We further propose **ERM-MinMaxGAP**, a fairness-aware training objective that augments empirical risk minimization with a penalty on the maximum within-language, within-modality male-female loss gap. By explicitly penalizing the worst subgroup disparity, the objective aims to improve overall SER performance while reducing gender imbalance within each language and modality. In summary, our contributions are twofold: (1) we provide, to the best of our knowledge, the first dedicated benchmark of gender bias in multilingual multimodal speech LLMs for SER; and (2) we propose ERM-MinMaxGAP, a fairness-aware objective that minimizes overall SER loss while directly reducing the worst within-language, within-modality gender disparity. Project page: <https://github.com/zihaupang/ERM-MinMaxGAP>.

2. Methodology

To minimize the overall gender gap in multilingual multimodal speech LLMs while preserving overall SER performance, we propose ERM-MinMaxGAP, as illustrated in Figure 1.

2.1. ERM-Based Supervised Fine-Tuning

Since 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 standard supervised fine-tuning (SFT) with Low-Rank Adaptation (LoRA) [31] in this study. We use the cross-entropy loss

** indicates the corresponding author.

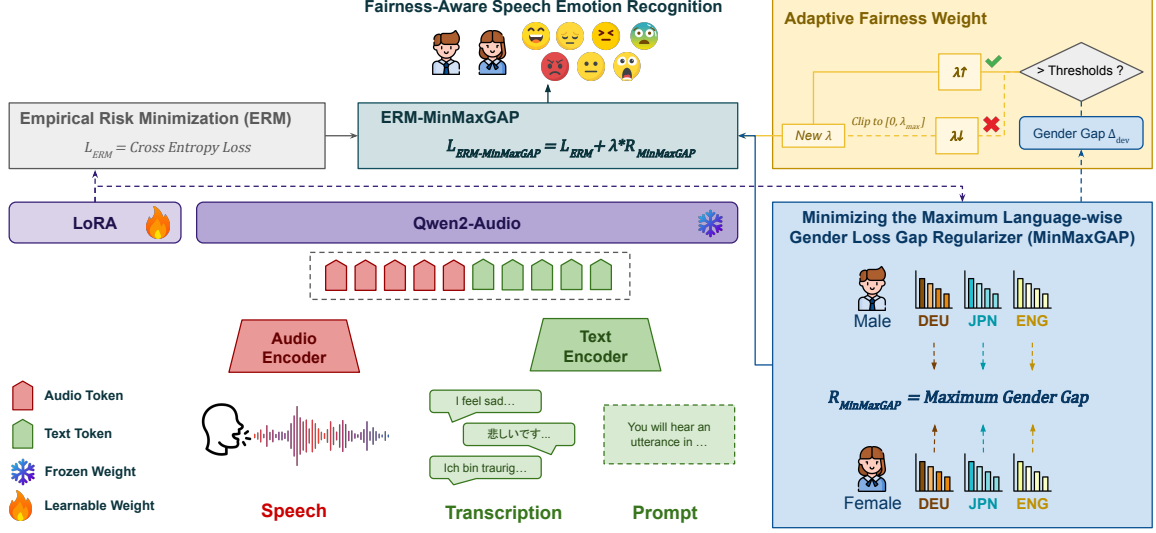


Figure 1: Architecture of the proposed method. The method consists of (1) empirical risk minimization for overall SER improvement, (2) MinMaxGAP for minimizing the language-wise gender gap, and (3) adaptive fairness-weight adjustment for fairness-aware SER.

[32] as our main objective function, denoted as $\mathcal{L}_{\text{ERM}}(\theta) = \text{CrossEntropyLoss}$, 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

$$\mathcal{L}_{\ell,g}(\theta) = \mathbb{E}[\text{CE}_i(\theta) \mid \ell_i = \ell, g_i = g],$$

where $\ell \in \mathcal{L}$ denotes the language, $g \in \{\text{F}, \text{M}\}$ denotes the gender group, and $\text{CE}_i(\theta)$ denotes the per-sample cross-entropy loss. The within-language gender gap for language ℓ is then defined as

$$\Delta_{\ell}(\theta) = |\mathcal{L}_{\ell,\text{F}}(\theta) - \mathcal{L}_{\ell,\text{M}}(\theta)|.$$

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

$$\mathcal{R}_{\text{MinMaxGAP}}(\theta) = (\Delta_{\max}(\theta))^p,$$

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.

In practice, this regularizer is computed using mini-batch estimates of the group-wise losses. For each batch, we compute the female and male losses within each language, form the corresponding language-wise gaps, and use the maximum valid gap as the batch-level fairness signal.

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 [33], we employ a Lagrange multiplier method to refine the weight adaptively during training (Figure 1 top right). We update the fairness weight as

$$\lambda^{(k+1)} = \Pi_{[0, \lambda_{\max}]} \left(\lambda^{(k)} + \eta \left(\Delta_{\text{dev}}^{(k)} - \epsilon \right) \right),$$

where $\Delta_{\text{dev}}^{(k)}$ 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 is already below the threshold, the update becomes small 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

$$\mathcal{L}_{\text{ERM-MinMaxGAP}}^{(k)}(\theta) = \mathcal{L}_{\text{ERM}}(\theta) + \lambda^{(k)} \mathcal{R}_{\text{MinMaxGAP}}(\theta).$$

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.

Table 1: Statistics of the MELD-ST dataset.

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

3. Experimental Setup

3.1. Database

We conduct experiments on **MELD-ST** [34], a multilingual emotion-aware speech dataset derived from MELD [35]. It extends the original English dialogue data with aligned English–Japanese and English–German speech pairs while preserving 7-class emotion labels. To support gender-fairness analysis, we additionally annotate speaker gender. Because the released data identify speakers but do not provide explicit gender labels, we manually assign one label per speaker by cross-checking video clips, audio, and cast information. We exclude utterances with multiple speakers or unreliable gender attribution. Each remaining utterance is associated with a single speaker and a unique gender label. Final dataset statistics are shown in Table 1.

3.2. Training Hyperparameters

We use **Qwen2-Audio-7B-Instruct** [14] as the backbone to fine-tune on the MELD-ST. We train for up to 20 epochs with an effective batch size of 64, learning rate 5×10^{-5} , weight decay 0.01, early stopping patience of 5, and random seed 42. For LoRA-based fine-tuning [31], we set $r = 16$, $\alpha = 32$, and dropout to 0.05. For the fairness objective, we use penalty power $p = 2$, initialize $\lambda = 0$, and enable adaptive updating with $\epsilon = 0.02$, update rate 0.5, and maximum $\lambda = 10.0$.

3.3. Comparison Models

We evaluate both **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 [14], Voxtral-Mini-3B [36], gpt4o-mini-audio¹, Kimi-Audio-7B-Instruct [37], and Ultravox-0.4².

3.4. Evaluation Metrics

We evaluate both **SER performance** and **gender fairness** [29, 38]. 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

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

²<https://github.com/fixie-ai/ultravox>

further report **AVG**, the mean of the four gap metrics, as a compact summary of overall gender bias.

4. Results and Analysis

4.1. 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., Voxtral-Mini-3B and 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.

4.2. 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.

4.3. 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; therefore, this setting was excluded from the study.

We also evaluated the effect of the fairness weight λ . Compared with our proposed adaptive weight update method, a fixed

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↓
<i>Multilingual</i>														
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 <i>↓1.07</i>
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 <i>↓2.25</i>
gpt4o-mini-audio	<u>45.89</u>	<u>44.57</u>	9.61	1.48	<u>2.61</u>	4.03	4.43	<u>52.65</u>	51.76	8.09	2.08	4.38	5.25	4.95 <i>↑0.52</i>
kimi-audio-7b	40.27	39.96	9.18	3.12	5.21	3.58	5.27	42.34	42.56	<u>7.15</u>	2.36	3.21	3.26	4.00 <i>↓1.27</i>
Ultravox-0.4	27.43	25.29	2.76	<u>1.49</u>	2.41	1.09	1.94	32.45	30.78	7.79	1.86	5.65	4.42	4.93 <i>↑2.99</i>
ERM-MinMaxGAP (Ours)	51.38	54.32	<u>6.38</u>	3.01	3.52	4.44	<u>4.34</u>	57.68	58.65	7.08	2.69	1.84	<u>2.53</u>	<u>3.53</u> <i>↓0.80</i>
<i>Monolingual</i>														
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 <i>↓2.15</i>
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 <i>↓2.71</i>
gpt4o-mini-audio	46.55	44.54	8.59	1.66	1.60	3.03	3.72	<u>52.40</u>	51.20	10.13	1.98	5.39	6.37	5.97 <i>↑2.25</i>
kimi-audio-7b	39.03	38.43	10.75	3.04	4.27	<u>2.13</u>	5.05	44.16	42.69	4.14	1.67	4.53	5.15	3.87 <i>↓1.17</i>
Ultravox-0.4	27.97	25.56	1.68	0.50	<u>1.92</u>	0.43	1.13	31.71	28.89	<u>4.97</u>	0.67	0.46	1.61	1.93 <i>↑0.79</i>
ERM-MinMaxGAP (Ours)	47.84	51.39	<u>3.06</u>	3.71	7.47	9.19	5.86	53.32	54.91	8.01	3.36	3.92	4.92	5.05 <i>↓0.81</i>
ENG														
Qwen2-Audio	46.02	44.27	7.69	3.65	<u>0.92</u>	1.75	3.50	46.13	41.67	10.34	2.31	4.57	5.11	5.58 <i>↑2.08</i>
Voxtral-Mini-3B	41.98	38.75	9.88	2.75	<u>7.57</u>	7.87	7.02	53.74	<u>57.71</u>	7.64	2.04	3.57	2.46	<u>3.93</u> <i>↓3.09</i>
gpt4o-mini-audio	49.67	49.06	12.47	1.52	4.22	5.93	6.03	<u>56.99</u>	56.35	<u>5.77</u>	1.36	4.42	5.29	4.21 <i>↓1.82</i>
kimi-audio-7b	<u>57.86</u>	<u>54.79</u>	10.33	1.54	5.52	5.63	5.80	56.11	53.85	10.58	<u>1.71</u>	<u>2.02</u>	2.17	4.77 <i>↓0.80</i>
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 <i>↑4.74</i>
ERM-MinMaxGAP (Ours)	58.09	60.52	<u>8.84</u>	2.10	0.41	<u>0.92</u>	<u>3.07</u>	68.13	68.85	5.08	1.74	0.38	1.24	2.11 <i>↓0.96</i>
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 <i>↓2.96</i>
Voxtral-Mini-3B	46.63	49.17	8.27	2.96	0.20	1.39	3.21	47.91	54.06	7.01	<u>2.12</u>	0.81	2.03	2.99 <i>↓0.22</i>
gpt4o-mini-audio	41.45	40.10	6.89	<u>1.24</u>	0.11	2.12	<u>2.59</u>	<u>48.56</u>	47.71	7.77	2.69	2.99	3.75	4.30 <i>↑1.71</i>
kimi-audio-7b	23.91	26.67	<u>5.50</u>	4.18	5.72	<u>1.50</u>	4.23	26.76	31.15	<u>4.95</u>	3.32	2.52	0.82	2.90 <i>↓1.32</i>
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 <i>↑2.16</i>
ERM-MinMaxGAP (Ours)	48.19	51.04	7.23	3.21	2.68	3.22	4.09	51.58	<u>52.19</u>	8.13	2.97	<u>1.22</u>	<u>1.44</u>	3.44 <i>↓0.64</i>

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 <i>↓1.07</i>
+ 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 <i>↓1.55</i>
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 <i>↓0.80</i>
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 <i>↓1.55</i>
$\lambda = 1$	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 <i>↓0.80</i>
$\lambda = 5$	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 <i>↑0.12</i>
$\lambda = 10$	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 <i>↑0.05</i>
$\lambda = \text{adaptive}$ (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 <i>↓0.80</i>
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 <i>↓0.61</i>
$p = 2$ (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 <i>↓0.80</i>

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 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.

6. Generative AI Use Disclosure

Generative AI tools were used only for language editing and polishing to improve the clarity and readability of this manuscript. They were not used to generate scientific content, research ideas, experimental designs, results, or conclusions. All content was reviewed by the authors, who take full responsibility for the manuscript.

7. Acknowledgment

This research was supported by Japan-Singapore Joint Call: Japan Science and Technology Agency (JST) (No. 251043539) and A*STAR 2024 (R24I6IR136), the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-004), and the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG Award No: AISG-NMLP-2024-003). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore.

8. References

- [1] B. W. Schuller, "Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends," *Communications of the ACM*, vol. 61, no. 5, pp. 90–99, 2018.
- [2] R. Sanjeewa, R. Iyer, P. Apputhurai, N. Wickramasinghe, and D. Meyer, "Empathic conversational agent platform designs and their evaluation in the context of mental health: systematic review," *JMIR Mental Health*, vol. 11, p. e58974, 2024.
- [3] S. Petrovica, A. Anohina-Naumeca, and H. K. Ekenel, "Emotion recognition in affective tutoring systems: Collection of ground-truth data," *Procedia Computer Science*, vol. 104, pp. 437–444, 2017.
- [4] J. M. Martín-Doñas, A. L. Zorrilla, M. deVelasco, J. C. Vázquez-Correa, A. Álvarez, M. I. Torres, P. Delgado, A. Lazpiur, B. Romero, and I. Alkorta, "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 (IC-NLSP 2024)*, 2024, pp. 119–128.
- [5] E. Jordan, R. Terrisse, V. Lucarini, M. Alrahabi, M.-O. Krebs, J. Desclés, and C. Lemey, "Speech emotion recognition in mental health: Systematic review of voice-based applications," *JMIR mental health*, vol. 12, no. 1, p. e74260, 2025.
- [6] Z. H. Pang, Y. Fu, Y. Gao, and T. Kawahara, "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)*. IEEE, 2026, pp. 19417–19421.
- [7] 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.
- [8] C. Busso, S. Parthasarathy, A. Burmania, M. AbdelWahab, N. Sadoughi, and E. M. Provost, "Msp-improv: An acted corpus of dyadic interactions to study emotion perception," *IEEE transactions on affective computing*, vol. 8, no. 1, pp. 67–80, 2016.
- [9] L. Pepino, P. Riera, and L. Ferrer, "Emotion Recognition from Speech Using wav2vec 2.0 Embeddings," in *Interspeech 2021*, 2021, pp. 3400–3404.
- [10] Y. Gao, C. Chu, and T. Kawahara, "Two-stage Finetuning of Wav2vec 2.0 for Speech Emotion Recognition with ASR and Gender Pretraining," in *Interspeech 2023*, 2023, pp. 3637–3641.
- [11] D. Diatlova, A. Udalov, V. Shutov, and E. Spirin, "Adapting wavlm for speech emotion recognition," *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024.
- [12] M. Sharma, "Multi-lingual multi-task speech emotion recognition using wav2vec 2.0," in *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2022, pp. 6907–6911.
- [13] 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," *The Twelfth International Conference on Learning Representations*, 2024.
- [14] 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.
- [15] P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov *et al.*, "Audiopalm: A large language model that can speak and listen," *arXiv preprint arXiv:2306.12925*, 2023.
- [16] D. Zhang, S. Li, X. Zhang, J. Zhan, P. Wang, Y. Zhou, and X. Qiu, "SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities," in *Findings of the Association for Computational Linguistics: EMNLP 2023*, H. Bouamor, J. Pino, and K. Bali, Eds. Singapore: Association for Computational Linguistics, Dec. 2023, pp. 15 757–15 773.
- [17] J. Bellver Soler, I. Martín Fernández, J. M. Bravo Pacheco, S. Esteban Romero, F. Fernández Martínez, and L. F. D'Haro Enríquez, "Multimodal audio-language model for speech emotion recognition," *The Speaker and Language Recognition Workshop (Odyssey 2024)*, 2024.
- [18] V. Sethu, J. Epps, and E. Ambikairajah, "Speaker variability in speech based emotion models-analysis and normalisation," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2013, pp. 7522–7526.
- [19] S. Mariooryad and C. Busso, "Compensating for speaker or lexical variabilities in speech for emotion recognition," *Speech Communication*, vol. 57, pp. 1–12, 2014.
- [20] J. Tavernor, Y. El-Tawil, and E. M. Provost, "The whole is bigger than the sum of its parts: Modeling individual annotators to capture emotional variability," *Interspeech 2024*, 2024.
- [21] M. A. Pastor, A. Ortega, and D. Ribas, "Analysis of the domain mismatch problem in the speech emotion recognition task," in *Proc. IberSPEECH 2024*, 2024, pp. 181–185.
- [22] I. R. Ulgen, Z. Du, C. Busso, and B. Sisman, "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)*. IEEE, 2024, pp. 12 081–12 085.
- [23] A. Koenecke, A. Nam, E. Lake, J. Nudell, M. Quartey, Z. Mengesha, C. Touns, J. R. Rickford, D. Jurafsky, and S. Goel, "Racial disparities in automated speech recognition," *Proceedings of the national academy of sciences*, vol. 117, no. 14, pp. 7684–7689, 2020.
- [24] R. Tatman, "Gender and dialect bias in YouTube's automatic captions," in *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, D. Hovy, S. Spruit, M. Mitchell, E. M. Bender, M. Strube, and H. Wallach, Eds. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 53–59.
- [25] C. Harris, C. Mgbahurike, N. Kumar, and D. Yang, "Modeling gender and dialect bias in automatic speech recognition," in *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024, pp. 15 166–15 184.
- [26] G. Attanasio, B. Savoldi, D. Fucci, and D. Hovy, "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*, 2024, pp. 21 318–21 340.

- [27] I.-E. Veliche, Z. Huang, V. A. Kochaniyan, F. Peng, O. Kalinli, and M. L. Seltzer, "Towards measuring fairness in speech recognition: Fair-speech dataset," *Interspeech 2024*, 2024.
- [28] C. Gorrostieta, R. Lotfian, K. Taylor, R. Brutti, and J. Kane, "Gender de-biasing in speech emotion recognition." in *Interspeech*, 2019, pp. 2823–2827.
- [29] Y.-C. Lin, H.-C. Chou, Y.-H. L. Liang, and H.-y. Lee, "Emo-debias: Benchmarking gender debiasing techniques in multi-label speech emotion recognition," *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU 2025)*, 2025.
- [30] Y.-C. Lin, T.-Q. Lin, C.-K. Yang, K.-H. Lu, W.-C. Chen, C.-Y. Kuan, and H.-y. Lee, "Listen and speak fairly: a study on semantic gender bias in speech integrated large language models," in *2024 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2024, pp. 439–446.
- [31] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen *et al.*, "Lora: Low-rank adaptation of large language models." *Iclr*, vol. 1, no. 2, p. 3, 2022.
- [32] Z. Zhang and M. Sabuncu, "Generalized cross entropy loss for training deep neural networks with noisy labels," *Advances in neural information processing systems*, vol. 31, 2018.
- [33] D. P. Bertsekas, *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.
- [34] S. Chen, S. Yahata, S. Shimizu, Z. Yang, Y. Li, C. Chu, and S. Kurohashi, "Meld-st: An emotion-aware speech translation dataset," in *Findings of the Association for Computational Linguistics: ACL 2024*, 2024, pp. 10 118–10 126.
- [35] 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.
- [36] A. H. Liu, A. Ehrenberg, A. Lo, C. Denoix, C. Barreau, G. Lample, J.-M. Delignon, K. R. Chandu, P. von Platen, P. R. Mudireddy *et al.*, "Voxtral," *arXiv preprint arXiv:2507.13264*, 2025.
- [37] K. Team, "Kimi-audio technical report," 2024.
- [38] Y.-C. Lin, H. Wu, H.-C. Chou, C.-C. Lee, and H.-y. Lee, "Emo-bias: A large scale evaluation of social bias on speech emotion recognition," *INTERSPEECH 2024*, 2024.