MOS-Bias: From Hidden Gender Bias to Gender-Aware Speech Quality Assessment

Wenze Ren [1,*], *Yi-Cheng Lin* [1,*], *Wen-Chin Huang* [2], *Erica Cooper* [3], *Ryandhimas E. Zezario* [4], *Hsin-Min Wang* [4], *Hung-yi Lee* [1], *Yu Tsao* [4,**]

[1] National Taiwan University, Taiwan
[2] Nagoya University, Japan
[3] National Institute of Information and Communications Technology, Japan
[4] Academia Sinica, Taiwan

{d14945014, f12942075, hungyilee}@ntu.edu.tw, yu.tsao@citi.sinica.edu.tw

Abstract

The Mean Opinion Score (MOS) serves as the standard metric for speech quality assessment, yet biases in human annotations remain underexplored. We conduct the first systematic analysis of gender bias in MOS, revealing that male listeners consistently assign higher scores than female listeners, a gap that is most pronounced in low-quality speech and gradually diminishes as quality improves. This quality-dependent structure proves difficult to eliminate through simple calibration. We further demonstrate that automated MOS models trained on aggregated labels exhibit predictions skewed toward male standards of perception. To address this, we propose a gender-aware model that learns gender-specific scoring patterns through abstract binary group embeddings, thereby improving overall and gender-specific prediction accuracy. This study establishes that gender bias in MOS constitutes a systematic, learnable pattern demanding attention in equitable speech evaluation.

Index Terms: speech quality assessment, mean opinion score, gender bias, listener perception

1. Introduction

The Mean Opinion Score (MOS), derived from human listening tests, is the gold standard for subjective speech quality assessment in tasks such as Text-to-Speech (TTS) [1], Voice Conversion (VC) [2], and Speech Enhancement (SE) [3]. However, large-scale listening tests are costly and time-consuming, driving the development of automated MOS prediction models, such as MOSNet [4], SSL-MOS [5], MOSA-Net [6], UTMOS [7, 8], and HighRateMOS [9], which estimate speech quality directly from audio. These models have become indispensable tools for speech quality assessment in modern speech generation systems [10].

Despite significant progress in improving the accuracy of automated MOS predictions, the quality of the MOS labels themselves has received limited attention [11, 12]. While biases in subjective listening tests are well-documented [13, 14], a critical yet overlooked source lies in the demographic composition of the listener pool. In particular, whether listener gender constitutes a systematic source of such bias remains largely unexplored. When male and female listeners hold systematically different perception standards [15], simple averaging produces a composite score that fails to accurately reflect either group and may even inadvertently favor one gender's perception standards while neglecting the other's.

In this study, we systematically examine for the first time how listener gender influences MOS ratings. Using the BVCC dataset [16], we reveal a consistent pattern: male listeners systematically assign higher scores, with this difference being most pronounced in low-quality speech and gradually diminishing as quality improves. However, standard MOS labels are derived by averaging all listener ratings without demographic calibration. This approach obscures meaningful differences in gender perception, and the resulting composite scores fail to accurately represent either group. Yet such averaged labels are widely regarded as gender-neutral benchmarks. Our empirical research challenges this assumption. We found that automatic MOS prediction models trained solely on these average labels (without any gender information) consistently predicted scores closer to male listeners' ratings than female listeners', even though each audio sample in the training set had fewer male listeners than female listeners on average. This reveals that average MOS labels are not gender-neutral: they implicitly encode male-perceived standards that models faithfully learn and propagate. To address this issue, we propose a gender-aware MOS prediction model that incorporates a parallel gender prediction branch. Rather than directly using listener gender labels as input, this model performs conditional processing based on abstracted binary group embeddings, enabling it to autonomously discover gender-specific perception patterns within the data. This approach enhances prediction accuracy both at the gender-specific and overall levels.

Our main contributions are summarized as follows:

1. We provide the first systematic evidence that male listeners consistently assign higher MOS scores than female listeners, with the gap largest in low-quality speech.
2. We show that averaged MOS labels and models trained on them implicitly inherit a male-leaning perceptual bias.
3. We propose a gender-aware model via abstract binary group embeddings, improving both overall and gender-specific prediction accuracy.

2. Related Work

2.1. Bias in Human Annotation

The impact of annotator demographics on annotation results has been extensively studied in Natural Language Processing (NLP) and Computer Vision (CV). In NLP, research has shown that annotators' gender, age, and cultural background introduce

[*] These authors contributed equally.
[**] indicates the corresponding author.

systematic biases in tasks such as sentiment analysis [17], hate speech detection, and toxicity labeling [18]. Similarly, studies in CV have demonstrated that annotators' demographic attributes affect the consistency of image annotations and the perception of offensive visual content [19, 20]. The broader literature on machine learning fairness further confirms that simple averaging of annotations tends to mask substantial discrepancies across demographic subgroups [21], leading to labels that overreact to the majority's views [22].

Although demographic factors, such as listeners' expertise and linguistic background, have been considered in speech quality assessment in the Blizzard Challenge [23, 24], these studies remain fragmented and lack a systematic approach. Notably, gender-related factors have been overlooked in speech assessment, despite strong evidence that gender influences auditory perception and evaluation behavior.

2.2. Fairness in Speech Processing

While fairness in speech processing has been extensively studied in tasks such as Speech Recognition (SR) [25, 26] and Speech Emotion Recognition (SER) [27, 28, 29], existing work predominantly takes the speaker's perspective. In SER specifically, Chien and Lee [30] demonstrated that male and female raters produce systematically different emotion annotations and proposed gender-specific perceptual models, while Shen et al. [31] personalized SER to individual annotators via meta-learning. In speech quality assessment, listener-dependent methods [11, 12] model individual listener tendencies, but do not examine their demographic structure.

3. Bias Analysis in MOS Annotations

3.1. Dataset and Toolkit

We adopted the SHEET [32] toolkit as our experimental framework, which supports MOS training and prediction across multiple public datasets. Among all datasets supported by SHEET, BVCC [16] is the only one providing both speaker and listener gender metadata, making it the only viable choice for our analysis. BVCC integrates publicly released samples from the Blizzard Challenge, Voice Conversion Challenge, and ESPnet-TTS, with each speech sample evaluated by eight listeners. The dataset is divided into 4,974 training speech samples, 1,066 development speech samples, and 1,066 test speech samples.

3.2. Problem Formulation

Let r_i denote the rating assigned by listener i to utterance u, and let $\mathcal{M}$ and $\mathcal{F}$ denote the sets of male and female listeners, respectively. The standard MOS is computed as:

$$\text{MOS}(u) = \frac{1}{N} \sum_{i=1}^{N} r_i \tag{1}$$

We define gender-specific scores as:

$$\text{MOS}_M(u) = \frac{1}{|\mathcal{M}|} \sum_{i \in \mathcal{M}} r_i, \quad \text{MOS}_F(u) = \frac{1}{|\mathcal{F}|} \sum_{i \in \mathcal{F}} r_i \tag{2}$$

The standard MOS is a weighted average of $\text{MOS}_M(u)$ and $\text{MOS}_F(u)$, with weights proportional to group size. If male and female listeners systematically differ in their ratings, this weighted average is closer to whichever group contributes more listeners. A single aggregated score, therefore, does not accurately represent the perceptual standards of either group. We quantify these differences in Sections 3.3 and 3.4, and propose a model that explicitly accounts for them in Section 5.

Table 1: *The mean and standard deviation of MOS scores for listeners of different genders corresponding to speakers of different genders in the BVCC training set.*

Listener	Speaker	Mean	Std	Count
Male	Male	2.925	1.137	9997
Male	Female	3.065	1.205	8229
Female	Male	2.822	1.175	11541
Female	Female	2.964	1.271	9598

Table 2: *Welch's t-test comparing MOS ratings between male and female listeners on the BVCC dataset. All differences are statistically significant ($p < 0.001$), indicating that male listeners consistently assign higher ratings than female listeners regardless of speaker gender.*

Condition	Male Listener	Female Listener	p
Male speaker	2.925	2.822	8.86×10^{-11}
Female speaker	3.065	2.964	5.35×10^{-8}
Overall	2.988	2.886	4.58×10^{-17}

3.3. Overall Gender Rating Difference

We quantified gender-based scoring differences within the training set. As shown in Table 1, male listeners consistently awarded higher scores than female listeners across all conditions: for male speakers, male listeners scored 2.925, while female listeners scored 2.822; for female speakers, the corresponding values were 3.065 and 2.964. This gap remained stable regardless of speaker gender. Notably, both listener groups rated female speakers higher, yet the pattern of "male listeners rating higher" persisted across all speaker gender conditions. This demonstrates that listener gender and speaker gender have independent influences on perceived quality.

To verify that this pattern was not due to sampling variation, we employed Welch's t-test, which is specifically designed to handle unbalanced sample sizes. As shown in Table 2, the scoring differences between male and female listeners were statistically significant across all conditions, with all p-values falling well below conventional significance thresholds. This confirms that the observed gap reflects genuine structural differences rather than random variation.

It is worth noting that this finding cannot be attributed to gender imbalance within the listener group. In the training set, each audio clip received ratings from 3.6 male listeners and 4.25 female listeners. Despite this, the rating disparity persisted, further confirming its structural nature.

3.4. Quality-Dependent Gender Difference

To examine how the gender gap varies with speech quality, we divided the training set into four quality intervals based on overall MOS: 1–2 (Poor), 2–3 (Fair), 3–4 (Good), and 4–5 (Excellent). For each interval, we computed the mean rating difference (male minus female) across all clips, stratified by speaker gender. Results are shown in Figure 1.

Averaged across speaker genders, the gap narrows monotonically with quality: 0.167 (1–2), 0.124 (2–3), 0.097 (3–4), and 0.030 (4–5). This pattern holds for each speaker gen-

Figure 1: *Across the four quality tiers of the BVCC training set, stratified by speaker gender, the mean rating difference between male and female listeners.*

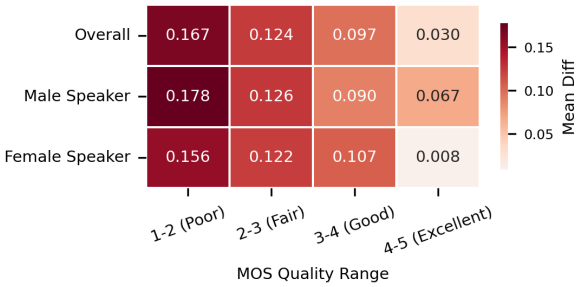


Table 3: *Bias inheritance analysis on the testing set. Model predictions are evaluated against ground truth (GT) from all (All), male (M), and female (F) listeners.*

GT	Utterance-Level				System-Level			
	LCC	SRCC	MSE	KTAU	LCC	SRCC	MSE	KTAU
All	**0.853**	**0.855**	**0.290**	**0.673**	**0.919**	**0.919**	**0.128**	**0.758**
M	0.806	0.811	0.372	0.628	0.901	0.899	0.141	0.727
F	0.802	0.801	0.430	0.614	0.888	0.886	0.194	0.713

der, indicating that the quality-dependent gap is driven by listener gender and is not an artifact of speaker gender. The reduced gap at the highest quality level is consistent with a ceiling effect, though greater perceptual agreement on high-quality speech may also play a role.

This quality-dependent structure means that gender differences are not a fixed offset: a single global correction cannot adequately capture a gap whose magnitude varies systematically with speech quality. This motivates models that can learn gender-specific scoring patterns across quality conditions, which we explore in Sec. 5.

4. Limitations of Gender-Agnostic MOS

We examine whether automated MOS prediction models, trained on these aggregated labels, produce predictions that differentially align with one gender group's ratings over the other.

4.1. Experimental Setup

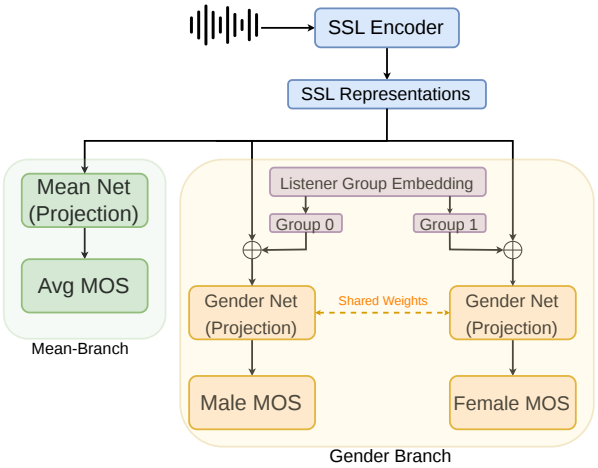
We adopt SSL-MOS as the baseline model, as it represents the mainstream framework for automated MOS prediction and serves as the foundational architecture supporting follow-up models (including UTMOS) by combining additional features.

To ensure robustness, we trained the model under three independent random seeds (1337, 2337, 3337) with all other conditions held constant. The model was optimized using SGD ($\text{lr} = 1 \times 10^{-3}$) for up to 100,000 steps, with early stopping triggered after 20 consecutive evaluations without improvement in system-level SRCC.

4.2. Bias Inheritance Analysis

We evaluate predictions at both the utterance level and the system level. We report Linear Correlation Coefficient (LCC) and Spearman Rank Correlation Coefficient (SRCC) for correlation, Mean Squared Error (MSE) for prediction error, and Kendall's

Figure 2: *The proposed gender-aware MOS prediction architecture. A shared SSL encoder feeds into two networks: a Mean Net predicting overall MOS, and a Gender Net predicting gender-specific MOS scores with shared projection weights.*



tau (KTAU) for ranking consistency. Table 3 presents results evaluated against three ground-truth references: overall MOS (All), male listener MOS (M), and female listener MOS (F), with all metrics averaged across three random seeds.

The model achieves its lowest error against the overall MOS it was trained on. When the same predictions are evaluated against gender-specific ground truths, we observe a consistent asymmetry: predictions are systematically closer to male listener ratings than to female listener ratings. Utterance-level MSE is 0.372 against male ground truth versus 0.430 against female ground truth (15.6% relative gap); at the system level, MSE is 0.141 versus 0.194 (37.6% relative gap). System-level LCC likewise favors male listeners (0.901 vs. 0.888).

Since the model receives no gender information during training, and its sole optimization target is the aggregated MOS, this asymmetry is notable: despite female listeners being the majority in the listener pool, model predictions consistently align more closely with male listener ratings. The prediction gaps in Table 3 empirically demonstrate a disparity that standard overall-only evaluation would not reveal. This motivates training objectives that explicitly account for gender-specific perception, which we explore in Sec. 5.

5. Gender-Aware MOS Prediction

We have demonstrated that gender bias exists in both human annotations and models trained on them. We further pose a deeper question: Can modeling gender-specific MOS patterns improve prediction performance over a single-label model? To address this, we designed a gender-aware framework based on the SSL-MOS architecture and evaluated whether explicitly modeling gender-specific rating behaviors can improve prediction accuracy.

5.1. Model Architecture

We start with the baseline SSL-MOS model and introduce a parallel prediction branch to enable gender-differentiated MOS estimation. Specifically, as shown in Figure 2, we replicate the MOS prediction network to create a "gender-MOS" branch that operates in parallel with the original "mean-MOS" branch. This joint architecture is carefully designed: rather than train-

Table 4: *Performance comparison on the BVCC test set, grouped by Ground Truth listener set. All results are averaged over three independent seeds (1337, 2337, 3337).* ***Prediction*** *denotes the branch output used for evaluation. Bold indicates the best result within each group. Shaded rows highlight the proposed gender-specific branch predictions.*

Model	Prediction	Ground Truth	Utterance-Level				System-Level			
			LCC	SRCC	MSE	KTAU	LCC	SRCC	MSE	KTAU
Overall Prediction Quality (GT: All Listeners)										
Baseline	Avg	All	0.853±0.002	0.855±0.001	0.290±0.004	0.673±0.002	0.919±0.004	**0.919**±0.005	0.128±0.001	0.758±0.006
Gender-MOS	Avg	All	**0.862**±0.003	**0.862**±0.002	**0.239**±0.016	**0.683**±0.002	**0.921**±0.005	0.918±0.003	**0.114**±0.020	**0.761**±0.004
Gender-Specific Prediction Quality (GT: Male Listeners)										
Baseline	Avg	Male	0.806±0.002	0.811±0.001	0.372±0.004	0.628±0.002	0.901±0.001	0.899±0.003	0.141±0.001	0.727±0.003
Gender-MOS	Avg	Male	0.816±0.001	0.819±0.002	0.335±0.016	0.636±0.002	0.904±0.003	0.903±0.001	0.142±0.023	**0.737**±0.002
Gender-MOS	Male	Male	**0.817**±0.001	**0.819**±0.002	**0.332**±0.012	**0.636**±0.002	**0.905**±0.003	**0.903**±0.001	**0.134**±0.017	0.736±0.002
Gender-Specific Prediction Quality (GT: Female Listeners)										
Baseline	Avg	Female	0.802±0.003	0.801±0.003	0.430±0.005	0.614±0.003	0.888±0.005	0.886±0.006	0.194±0.001	0.713±0.009
Gender-MOS	Avg	Female	0.806±0.004	0.806±0.000	0.366±0.016	0.620±0.000	0.888±0.006	0.886±0.005	**0.165**±0.018	0.711±0.005
Gender-MOS	Female	Female	**0.807**±0.005	**0.807**±0.000	**0.366**±0.028	**0.621**±0.001	**0.888**±0.007	**0.887**±0.005	0.169±0.030	**0.713**±0.006

ing two completely independent models for male and female listeners (which would reduce the effective training data available to each), the shared SSL encoder enables both branches to benefit from the full dataset while still learning gender-specific rating patterns. Meanwhile, the key design decision is how to input listener gender information: we avoid directly concatenating gender labels (which would violate the baseline model's gender-neutral design) and instead condition the gender-MOS branch on two hard-coded binary values: 0 and 1. These values represent two abstract listener groups that the model must learn autonomously: Group 1 corresponds to male rating patterns, and Group 0 corresponds to female rating patterns. This design embodies a deliberate stance: the base model operates as a gender-agnostic listener, and we expect it to naturally develop gender perception through exposure to real-world gendered data signals (rather than explicit demographic labels). The final architecture produces three parallel outputs: the average MOS prediction for the original branch, the MOS prediction from the male listener's perspective for the gender MOS branch, and the MOS prediction from the female listener's perspective.

5.2. Training Objectives

We define the training objective as a multi-task loss composed of three weighted components:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{avg}} + \mathcal{L}_{\text{male}} + \mathcal{L}_{\text{female}}$$

where each term is the MSE between the corresponding branch prediction and its gender-specific ground truth. Equal weighting (1:1:1) ensures no single perspective dominates the optimization.

5.3. Results and Analysis

Overall Performance. As shown in Table 4, when the ground truth is all listeners, the gender-aware model achieves better overall prediction quality than the baseline. At the utterance level, the gender-aware model achieves an LCC of 0.862 and an MSE of 0.239, while the baseline model achieves 0.853 and 0.290, respectively. This suggests that the auxiliary specificity objective may provide complementary training signals that assist the primary prediction branch rather than compete with it.

Gender-Specific Accuracy. As shown in Table 4, when the ground truth is male or female listeners respectively, gender-specific branches improved prediction accuracy for each gender. For male listeners, the LCC for the utterance-set rose from 0.806 (baseline) to 0.817, and MSE dropped from 0.372 to 0.332. For female listeners, LCC increased from 0.802 to 0.807, and MSE fell from 0.430 to 0.366. The improvement for male listeners exceeded that for females, possibly reflecting stronger internal consistency in male ratings. Within the mean-branch of the gender MOS model, gender-specific prediction errors were also lower than in the baseline model (MSE decreased from 0.372 to 0.335 for males and from 0.430 to 0.366 for females). This shows that multi-task learning enhances both overall prediction accuracy and the modeling of gender pattern differences.

Results Analysis. These results yield three key conclusions: First, the model successfully learns gender-specific scoring patterns solely through abstract binary encoding, capturing the systematic differences between the two predefined listener groups. Second, explicitly modeling gender-related variance enables the primary prediction branch to learn intrinsic speech quality more purely. Third, the availability of an independent gender prediction head lays the practical groundwork for future fairness-oriented interventions.

6. Conclusion

This study presents the first systematic investigation of gender bias in MOS-based speech quality assessment. We find that male listeners consistently assign higher ratings than female listeners, with the gap largest for low-quality speech (mean difference of 0.167) and diminishing to near zero at excellent quality, a quality-dependent structure that fixed calibration strategies are unlikely to correct. We further show that this bias propagates into automated models: even without any gender supervision, model predictions align systematically closer to male ratings, with system-level MSE gaps reaching up to 37.6% relative to female ground truth. To address this, we propose a gender-aware model conditioned on abstract binary group embeddings. Notably, the model successfully discovers gender-specific perceptual patterns without explicit demographic labels, improving both overall and gender-specific prediction accuracy. Future work will develop bias mitigation methods for MOS labels and models, and validate our approach across additional datasets. We expect this research to spur the learning community to critically examine fairness issues in speech quality assessment and advance toward more equitable evaluation practices.

7. Generative AI Use Disclosure

Generative AI tools assisted in the linguistic polishing of the manuscript. The authors remain solely responsible for the research design, experiments, analysis, and reported results. AI tools did not contribute to the substantive scientific content.

8. References

[1] C.-J. Hsu, Y.-C. Lin, C.-C. Lin, W.-C. Chen, H. L. Chung, C.-A. Li, Y.-C. Chen, C.-Y. Yu, M.-J. Lee, C.-C. Chen *et al.*, "Breezyvoice: Adapting tts for taiwanese mandarin with enhanced polyphone disambiguation–challenges and insights," *arXiv preprint arXiv:2501.17790*, 2025.

[2] W.-C. Huang, H.-T. Hwang, Y.-H. Peng, Y. Tsao, and H.-M. Wang, "Voice conversion based on cross-domain features using variational auto encoders," in *Proc. ISCSLP 2018*.

[3] X. Lu, Y. Tsao, S. Matsuda, and C. Hori, "Speech enhancement based on deep denoising autoencoder." in *Proc. Interspeech 2013*.

[4] C.-C. Lo, S.-W. Fu, W.-C. Huang, X. Wang, J. Yamagishi, Y. Tsao, and H.-M. Wang, "MOSNet: Deep learning-based objective assessment for voice conversion," in *Proc. Interspeech 2019*.

[5] E. Cooper, W.-C. Huang, T. Toda, and J. Yamagishi, "Generalization ability of mos prediction networks," in *Proc. ICASSP 2022*.

[6] R. E. Zezario, S.-W. Fu, F. Chen, C.-S. Fuh, H.-M. Wang, and Y. Tsao, "Deep learning-based non-intrusive multi-objective speech assessment model with cross-domain features," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 54–70, 2022.

[7] T. Saeki, D. Xin, W. Nakata, T. Koriyama, S. Takamichi, and H. Saruwatari, "UTMOS: UTokyo-SaruLab system for VoiceMOS Challenge 2022," in *Pro. Interspeech 2022*.

[8] K. Baba, W. Nakata, Y. Saito, and H. Saruwatari, "The t05 system for the voicemos challenge 2024: Transfer learning from deep image classifier to naturalness mos prediction of high-quality synthetic speech," in *Proc. SLT 2024*.

[9] W. Ren, Y.-C. Lin, W.-C. Huang, R. E. Zezario, S.-W. Fu, S.-F. Huang, E. Cooper, H. Wu, H.-Y. Wei, H.-M. Wang, H.-Y. Lee, and Y. Tsao, "HighRateMOS: Sampling-rate aware modeling for speech quality assessment," in *Proc. ASRU 2025*.

[10] C.-K. Yang, Y.-K. Fu, C.-A. Li, Y.-C. Lin, Y.-X. Lin, W.-C. Chen, H. L. Chung, C.-Y. Kuan, W.-P. Huang, K.-H. Lu *et al.*, "Building a taiwanese mandarin spoken language model: A first attempt," *arXiv preprint arXiv:2411.07111*, 2024.

[11] Y. Leng, X. Tan, S. Zhao, F. Soong, X.-Y. Li, and T. Qin, "Mbnet: Mos prediction for synthesized speech with mean-bias network," in *Proc. ICASSP 2021*.

[12] W.-C. Huang, E. Cooper, J. Yamagishi, and T. Toda, "Ldnet: Unified listener dependent modeling in mos prediction for synthetic speech," in *Proc. ICASSP 2022*.

[13] S. Zielinski, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests-a review," *Journal of the Audio Engineering Society*, vol. 56, no. 6, pp. 427–451, 2008.

[14] A. Rosenberg and B. Ramabhadran, "Bias and Statistical Significance in Evaluating Speech Synthesis with Mean Opinion Scores," in *Proc. Interspeech 2017*.

[15] N. Aloufi, A. Heinrich, K. Marshall, and K. Kluk, "Sex differences and the effect of female sex hormones on auditory function: a systematic review," *Frontiers in Human Neuroscience*, vol. 17, p. 1077409, 2023.

[16] E. Cooper and J. Yamagishi, "How do voices from past speech synthesis challenges compare today?" in *Proc. 11th ISCA SSW, 2021*.

[17] H. Al Kuwatly, M. Wich, and G. Groh, "Identifying and measuring annotator bias based on annotators' demographic characteristics," in *Proc. WOAH 2020*.

[18] M. Sap, S. Swayamdipta, L. Vianna, X. Zhou, Y. Choi, and N. A. Smith, "Annotators with attitudes: How annotator beliefs and identities bias toxic language detection," in *Proc. NAACL 2022*.

[19] P. S. Sachdeva, R. Barreto, C. von Vacano, and C. J. Kennedy, "Assessing annotator identity sensitivity via item response theory: A case study in a hate speech corpus," in *Proc. FAccT 2022*.

[20] Z. Liao, S. Hu, Y. Xie, and Y. Xia, "Transformer-based annotation bias-aware medical image segmentation," in *Proc. MICCAI 2023*.

[21] A. Mostafazadeh Davani, M. Díaz, and V. Prabhakaran, "Dealing with disagreements: Looking beyond the majority vote in subjective annotations," *Transactions of the Association for Computational Linguistics*, vol. 10, pp. 92–110, 2022.

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

[23] X. Zhou, Z.-H. Ling, and S. King, "The blizzard challenge 2020." in *Proc. VCCBC 2020*.

[24] S. King and V. Karaiskos, "The blizzard challenge 2013," in *Proc. Blizzard 2013*.

[25] C. Harris, C. Mgbahurike, N. Kumar, and D. Yang, "Modeling gender and dialect bias in automatic speech recognition," in *Proc. Findings of EMNLP 2024*.

[26] H. ElGhazaly, B. Mirheidari, N. S. Moosavi, and H. Christensen, "Exploring gender disparities in automatic speech recognition technology," *arXiv preprint arXiv:2502.18434*, 2025.

[27] Y.-C. Lin *et al.*, "Mitigating subgroup disparities in multi-label speech emotion recognition: A pseudo-labeling and unsupervised learning approach," in *Proc. Interspeech 2025*.

[28] Y.-S. Tsai, Y.-C. Lin, H.-C. Chou, and H.-Y. Lee, "Co-vada: A confidence-oriented voice augmentation debiasing approach for fair speech emotion recognition," in *Proc. ASRU 2025*.

[29] Y.-C. Lin, H.-C. Chou, Y.-H. L. Liang, and H.-Y. Lee, "Emodebias: Benchmarking gender debiasing techniques in multi-label speech emotion recognition," in *Proc. ASRU 2025*.

[30] W.-S. Chien and C.-C. Lee, "Achieving fair speech emotion recognition via perceptual fairness," in *Prof. ICASSP 2023*.

[31] S.-X. Fang, L.-Y. Shen, Y.-C. Lin, H.-C. Chou, and H.-Y. Lee, "Meta-perser: Few-shot listener personalized speech emotion recognition via meta-learning," in *Proc. Interspeech 2025*.

[32] W.-C. Huang, E. Cooper, and T. Toda, "MOS-Bench: Benchmarking generalization abilities of subjective speech quality assessment models," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 34, pp. 2385–2397, 2026.