Justice in the Waveform: Synthesizing Group Fairness, Individual Consistency, and Cross-Corpus Robustness in Speech Emotion Recognition

Himani Thakor
Ph.D. Scholar,
Department of Computer Science
Veer Narmad South Gujarat University, Surat, Gujarat, India
E-Mail: ht3991214@gmail.com

Abstract:

Speech Emotion Recognition (SER) is increasingly deployed as an institutional gatekeeper in high-stakes environments, yet algorithmic fairness in speech AI remains fragmented. This paper synthesizes contemporary findings across ten foundational studies to address the multi-dimensional challenges of speech-based algorithmic bias. We analyze the core mathematical and structural tensions between group-level statistical parity and individual consistency in perceptually fair models. We examine the cross-corpus generalization gap, mapping how performance degrades by 10% to 40% under acoustic and stylistic mismatches, particularly when transitioning from acted laboratory recordings to spontaneous, real-life speech. Additionally, we evaluate data-centric interventions—including generative spectrogram augmentation via Generative Adversarial Networks (GANs) and ImageNet transfer learning—alongside computational scaling behaviors under parameterized resource constraints. By mapping these diverse technical paradigms, we provide a unified sociotechnical framework that links formal fairness definitions, pipeline-aware bias diagnostics, and efficient optimization strategies for equitable and robust speech technologies.

1. Introduction:

As voice-enabled systems evolve from supportive applications into institutional decision-makers, speech technologies are increasingly deployed in high-stakes sectors such as clinical diagnosis, automated employment interviewing, and emergency dispatch systems. This rapid deployment has initiated a critical sociotechnical turn in speech AI, recognizing that systems do not operate in a vacuum but are deeply embedded within and reflective of social hierarchies. When speech technologies fail, they do not produce random errors; instead, they often reproduce and entrench existing societal inequities, establishing service disparities for marginalized groups or enforcing rigid linguistic norms that treat diverse cultural identities as deviations to be corrected.

Addressing algorithmic bias in the speech domain presents physical and mathematical challenges unique to the auditory modality, termed acoustic entanglement. In textual models, sensitive variables (e.g., gendered pronouns) can be localized and removed at the discrete token level. In visual models, sensitive features (e.g., skin tone or face regions) can be masked or cropped. By contrast, the continuous speech signal jointly encodes lexical content, speaker identity, accent, age, gender, and affect across continuous, overlapping frequency bands. Naive attempts to strip demographic variables or enforce invariant representations inevitably degrade downstream task utility or erase essential pragmatic cues, leaving no direct analog to simple feature suppression or token deletion.

This paper synthesizes a comprehensive taxonomy of social bias and algorithmic fairness across contemporary speech AI research. We distinguish three temporal mechanisms of bias: (1) pre-existing bias, which is absorbed from standard language ideologies before system design; (2) technical bias, which arises from design choices such as legacy telephony codecs that degrade higher-pitched voices; and (3) emergent bias, which materializes post-deployment when systems encounter naturalistic, non-standard, or pathological speech. These mechanisms manifest either allocative harms (e.g., systematic service disparities in automated transcriptions) or representational harms (e.g., the reinforcing of gender stereotypes via submissive digital assistant voices, linguistic erasure of minority ethnolects, or misgendering through binary fundamental frequency classification). We trace these challenges across the model lifecycle, from data curation to efficient model scaling and downstream rater annotations.

2. Group Fairness versus Individual Consistency

A fundamental challenge in modern speech AI is the mathematical incompatibility between different definitions of fairness. Developers typically focus on group fairness, which seeks to achieve equal average outcomes across demographic groups. However, optimizing strictly for group-level statistical parity can severely degrade individual consistency, which requires that similar individual utterances receive similar predictions.

Chien and Lee investigated this core tension in perceptually fair SER systems by training binary emotion detectors (Neutral, Happiness, Anger, Sadness) using 768-dimensional latent wav2vec 2.0 embeddings. To eliminate gender-based group bias, they optimized a perceptually fair model with a multi-task loss function:

$$LTotal = LR - LAdv + \lambda LD \qquad (1)$$

where LR is the standard cross-entropy loss for emotional outcome prediction, LAdv evaluates how easily speaker gender can be detected from the learned embeddings (employing an adversarial discriminator to strip gender cues), and LD is a distance-based regularizer that minimizes the distance between gender classes in the latent feature space.

Their systematic analysis revealed a sharp tradeoff: as gender-based group information was progressively removed from the model's representations, the system's individual consistency (evaluated via a k-nearest neighbor metric with k = 20) steadily deteriorated. When more than 70% of the biased training data was utilized to eliminate gender information, individual fairness fell below the baseline, unmitigated model's scores. Crucially, the severity of this trade-off is modulated by two distinct parameters:

- Rater Diversity and Corpus Scale: On the highly curated IEMOCAP dataset, which features only 6 raters and a strong male consensus bias, the drop in individual consistency remained under 4%. On the larger, naturalistic BIIC-Podcast dataset, which incorporates emotional annotations from 89 diverse raters, enforcing group fairness led to a severe individual fairness drop of up to 20%. This highlights how coarse group parity masks the subjective, rich variations of individual expressions preserved in larger annotator panels.
- Wasserstein Distance (WD) Thresholds: When segmenting the dataset into quartiles based on the Wasserstein Distance between gender-subgroup attribute distributions in the embedding space, individual consistency remained stable for subsets below the 75th percentile of WD. However, for acoustically distant subsets (where the gender groups exhibit highly distinct formant and pitch structures), forcing demographic alignment led to a sharp, catastrophic decline in individual fairness.

3. The Spontaneity Gap and Cross-Corpus Fragility

A persistent barrier to practical SER deployment is the dramatic drop in model accuracy when tested on unseen acoustic conditions or speaking styles. Models optimized on a single corpus typically achieve high accuracy in matched conditions but exhibit a performance degradation of 10% to 40% in mismatched cross-corpus testing.

This fragility is primarily driven by the spontaneity gap. Speech corpora occupy a wide spectrum of authenticity: acted databases utilize trained actors reading static scripts, producing stereotypical and exaggerated emotions; elicited databases induce emotional states through designed interactive scenarios; while spontaneous databases capture genuine, natural conversations. To benchmark this gap, Gómez-Zaragozá et al. introduced the EMOVOME dataset, consisting of 999 authentic WhatsApp voice messages collected from 100 Spanish speakers in real-life, unguided settings (see Table 1). Benchmarking state-of-the-art self-supervised architectures on EMOVOME, IEMOCAP, and RAVDESS demonstrated that model performance scales directly with dataset controlled Ness:

1. **Acted Ceiling:** Pre-trained transformer models (e.g., UniSpeech-SAT-Large) combined with extended Geneva Minimalistic Acoustic Parameter Set (eGeMAPS) features achieved 73.54% unweighted accuracy (UA) for 3-class valence and 71.94% for arousal prediction on the acted RAVDESS corpus.
2. **Elicited Mid-point:** On the elicited, text-independent IEMOCAP corpus, the same architecture achieved 60.04% UA for valence and 61.20% for arousal.

3. **Spontaneous Floor:** On the spontaneous EMOVOME corpus, the performance fell to 57.53% UA for valence and 43.57% for arousal under expert-only labels, showing a massive degradation under real-life background noise, variable hardware, and non-stereotypical emotional deliveries.

Table 1: Taxonomy and Metadata Comparison of Core Speech Emotion Recognition Corpora

Dataset	Authenticity	Environment	Language	Speakers (M/F)	Rater Base	Mean Duration
EMOVOME	Spontaneous	Real-life (WhatsApp)	Spanish	100 (50 / 50)	2 Experts + 3 Non-experts	17.59s
IEMOCAP	Elicited / Acted	Laboratory Room	English	10 (5 / 5)	6 Experts (Consensus)	4.59s
RAVDESS	Acted	Studio Box	English	24 (12 / 12)	None (Designed Script)	3.70s
CREMA-D	Acted	Controlled Lab	English	91 (48 / 43)	2,443 Crowd-workers	2.50s

Furthermore, the EMOVOME study exposed critical dependencies on annotator expertise and demographics, directly influencing algorithmic fairness:

- The Expertise Disparity: Models trained strictly on expert annotations (provided by clinical psychologists) exhibited severe gender bias, displaying a 10% unweighted accuracy gap favoring male speakers. This indicates that experts can possess implicit gender-specific expectations regarding acoustic expressions of distress or affect.
- The Collaborative Solution: Incorporating non-expert crowdsourced labels and training on combined expert-non-expert labels reduced individual rating biases, raising unweighted accuracy while producing the most gender-equitable outcomes (achieving a marginal gender difference of only 0.3% for valence and 1.4% for arousal).

4. Data Efficiency and Synthetic Mitigation

Because large-scale, emotionally annotated datasets are difficult to collect and annotating spontaneous voice data is highly subjective, SER models are frequently constrained by extreme data scarcity and severe class imbalances (e.g., neutral speech samples vastly outnumbering highly charged emotional expressions like sadness or anger).

To resolve these data bottlenecks, contemporary research investigates two complementary paradigms: image-based transfer learning and generative spectrogram augmentation.

4.1 Cross-Domain Transfer Learning

Treating audio waveforms as 2D visual representations—specifically, log-mel spectrogram arrays—allows researchers to leverage the feature-extraction capabilities of models pre-trained on massive visual corpora. Tai Vu demonstrated that a standard ResNet34 architecture, pre-trained on ImageNet, can be successfully fine-tuned for SER on highly restricted datasets.

Using a combined dataset of RAVDESS and SAVEE, fine-tuning the ImageNet-pretrained ResNet34 achieved an accuracy of 66.7% and an F1 score of 0.631, representing a substantial improvement over models trained from scratch. This cross-domain transfer learning was reinforced by two critical data-centric techniques:

- Progressive Resizing: CNN models were first pre-trained on compact 128 × 128 log-mel spectrograms to accelerate initial convergence, then fine-tuned on larger 256 × 256 arrays, forcing the model to learn multi-scale acoustic features.
- Mixup Regularization: Generating convex combinations of randomly paired training spectrograms and their corresponding labels ($x_{mix} = \alpha x_i + (1 - \alpha)x_j$) effectively smoothed the decision boundaries and prevented severe overfitting on limited training samples.

4.2 Generative Augmentation via GANs

To address high class imbalance directly without performing loss-weighting or random under sampling, Chatziagapi et al. proposed using conditional Generative Adversarial Networks (GANs) to synthesize high-fidelity mel-scaled spectrograms for underrepresented minority classes. Adapting the Balancing GAN (BAGAN) framework, they introduced four key architectural modifications to ensure training stability and synthesize high-resolution spectrograms:

- Replacing all up-sampling layers in the generator with transposed convolutions to prevent visual artifacts and extreme pixel boundaries.
- Implementing Leaky ReLU activation across all intermediate layers.
- Utilizing batch normalization and dropout ($p = 0.2$) in both G and D.
- Feeding real and synthetic batches to the discriminator separately to prevent batch-normalization-induced leakage.

To accelerate convergence and prevent the common failure mode of mode collapse, they pre-trained the GAN's generator and discriminator using a stacked autoencoder trained on the

entire imbalanced corpus. Class conditioning was then initialized by calculating the mean and covariance matrix of the learned latent vectors to model each emotion class as a multivariate normal distribution. This generative approach achieved a 10% relative unweighted average recall (UAR) improvement on the IEMOCAP dataset and a 5% relative improvement on the highly imbalanced, multi-domain FEEL-25k corpus, significantly outperforming traditional signal-space augmentations (e.g., speed perturbation or pitch shifting).

5. Computational Scaling and Parameter Efficiency

The deployment of large-scale pre-trained speech models (such as Whisper or wav2vec 2.0) introduces immense computational burdens. Designing efficient systems requires understanding how to allocate fixed compute budgets across three fundamental axes: model size (xN), input duration (xT), and representational resolution (xV). Wu et al. conducted a joint study of compute constraints and optimization behavior, highlighting that scaling behavior is highly task-dependent. In automatic speech recognition (ASR), scaling is smooth and monotonic: increasing model size and doubling input context yields predictable, consistent reductions in Word Error Rate (WER). For SER, however, the compute-performance Pareto frontier is highly sparse and operates under distinct constraints:

- Model Size Diminishing Returns: Upgrading from a fine-tuned wav2vec2-base (95M parameters) to a wav2vec2-large-robust (317M parameters) provides minor, saturating improvements in emotional discriminability, suggesting that task-specific complexity saturates at moderate parameter capacities.
- The Temporal Sweet Spot: While ASR benefits from maximum input context, SER exhibits a non-trivial temporal optimum of exactly 4 seconds (evaluated on CREMA-D). Clips shorter than 4s truncate essential prosodic contours, while clips longer than 4s introduce padding artifacts and neutral speech content that degrades classification accuracy. Clipping audio inputs to 4s reduced total FLOPs by 33% while achieving a superior unweighted accuracy of 56.49% compared to 6s inputs (55.31%).
- The Adaptation Collapse: Standard parameter-efficient fine-tuning (PEFT) using LoRA-only adaptation collapsed to 43.82% UA on 6-class CREMA-D, which is barely above the random baseline of 16.7%. Achieving high-performance adaptation requires combining LoRA with depth-aware layer unfreezing (unfreezing the top 4 or top 8 transformer encoder layers), demonstrating that deep representations must be reshaped to align pre-trained acoustic features with complex emotional dimensions.

6. Discussion and Sociotechnical Outlook

Integrating these diverse research findings reveals that solving algorithmic bias in speech AI requires moving beyond simple performance parity metrics. Our synthesis outlines several open directions for the field:

1. **Standardized Metadata Protocols:** To combat data-related bias, research must transition toward structured, multi-dimensional metadata reporting. Datasets should document recording hardware configurations, environmental signal-to-noise ratios, and the demographic and cultural profiles of both speakers and annotators.
2. **Post-Training Homogenization:** Modern Large Audio Models (LAMs) optimized via Reinforcement Learning from Human Feedback (RLHF) or Direct Preference Optimization (DPO) face severe cultural homogenization. Post-training alignment encourages models to adopt narrow, standardized speaking styles rated highly by dominant annotator groups, systematically penalizing dialectal prosody, non-standard accents, and code-switching.
3. **Privacy-Preserving Auditing:** The enforcement of strict privacy and data sovereignty laws (e.g., GDPR) frequently starves fairness audits of the very demographic metadata required to evaluate bias. Developing high-fidelity synthetic voice conversion and demographic-preserving perturbation models represents a vital avenue toward a privacy-preserving fairness regime.

7. Conclusion

This paper has synthesized contemporary research in speech AI fairness, analyzing how acoustic entanglement, spontaneity, and annotator subjectivity compound along the model processing pipeline. By evaluating the trade-offs between group and individual fairness, mapping cross-corpus generalization gaps across acted and spontaneous speech, and analyzing the Pareto-optimal scaling boundaries under fixed compute constraints, we highlight that fairness in speech AI is not a post-hoc optimization heuristic. It is a multi-dimensional, sociotechnical practice of justice that must be integrated into every stage of voice-enabled system design, from participatory data collection to depth-aware, resource-efficient model adaptation.

References

1. W.-S. Chien and C.-C. Lee, “An Investigation of Group versus Individual Fairness in Perceptually Fair Speech Emotion Recognition,” in Proceedings of Interspeech, 2024, pp. 3205–3209.
2. Y.-C. Lin et al., “Toward Fair Speech Technologies: A Comprehensive Survey of Bias and Fairness in Speech AI,” arXiv preprint arXiv:2412.03511, 2024.

3. L. Gómez-Zaragozá et al., “EMOVOME: A Dataset for Emotion Recognition in Spontaneous Real-Life Speech,” arXiv preprint arXiv:2403.09990, 2024.
4. N. Braunschweiler, R. Doddipatla, S. Keizer, and S. Stoyanchev, “A Study on Cross-Corpus Speech Emotion Recognition and Data Augmentation,” in arXiv preprint arXiv:2201.03511, 2022.
5. A. Chatziagapi et al., “Data Augmentation using GANs for Speech Emotion Recognition,” in Proceedings of Interspeech, 2019, pp. 1218–1222.
6. T. Vu, “Amplifying Emotional Signals: Data-Efficient Deep Learning for Robust Speech Emotion Recognition,” arXiv preprint arXiv:2501.01802, 2025.
7. J. Wu et al., “Scaling Audio Models Efficiently: A Joint Study of Compute Constraints and Optimization Behavior,” arXiv preprint arXiv:2602.06922, 2026.
8. A. Ibrahim et al., “What Does it Take to Generalize SER Model Across Datasets? A Comprehensive Benchmark,” arXiv preprint arXiv:2603.01853, 2026.