Normalization in Speech Recognition, A Theoretical Algorithm to Improve Gender Bias Before We Use An ASR Model

Dr. Cecilia Alm
Program Director Artificial Intelligence
Department of Psychology
cecilia.o.alm@rit.edu

Steven Szachara
M.S. Data Science Student
Department of Data Science
ss9270@rit.edu

Andrew Murphy
M.S Artificial Intelligence Student
School of Information
acm7552@rit.edu

1 Introduction

Automatic Speech Recognition (ASR) systems have advanced significantly over recent decades, yet their performance is still highly sensitive to acoustic variability. Differences in speaker identity, speaking rate, recording environment, and background noise introduce systematic distortions that can alter the statistical distribution of speech features and degrade recognition accuracy. As a result, normalization has become a foundational strategy in ASR front-end processing, aiming to transform raw or derived speech features into a more stable and invariant representation before model inference. This review examines major approaches to quantifying ASR bias in subgroups of interest, bias mitigation strategies both data- and model-driven, as well as normalization strategies across prior literature with a focus on mathematical techniques applied directly to the speech feature domain. To combat gender bias in ASR, we propose **Gender-Responsive Adaptive Feature Normalization (GRAFN)**, a raw waveform–level normalization method applied before feature extraction or model training, and measure Word Error Rate (WER) parity on state-of-the-art ASR systems OpenAI's Whisper model [1] and Meta AI's wav2vec2 [2] finetuned on Mozilla Common Voice, both before and after applying GRAFN to the training dataset.

1.1 Normalization Background

Normalization methods in speech recognition generally aim to mitigate mismatches between training and testing conditions by adjusting either the statistical distribution or structural properties of feature representations. Early approaches focused on simple feature-space adjustments such as cepstral mean normalization (CMN) and mean-variance normalization (MVN), which assumed that stationary shifts in channel or environmental characteristics could be corrected through linear centering and scaling. More advanced techniques later emerged to compensate for deeper non-linearities, address temporal variability such as speaking rate, or explicitly account for inter-speaker physiological differences through spectral warping or affine transformations [3, 4, 5, 6, 7, 8]. These methods vary in terms of where they operate in the signal processing pipeline, the mathematical model used to define the normalization space, and the types of variability they target.

2 Mathematical Normalization in Speech Data

One category of normalization techniques operates directly on the feature vector by enforcing statistical consistency over a temporal window. For example, the recursive mean and variance normalization method computes a locally stationary normalization of Mel-Frequency Cepstral Coefficient (MFCC) features,

$$\hat{x} = \Lambda \left(x - \mu \right), \tag{1}$$

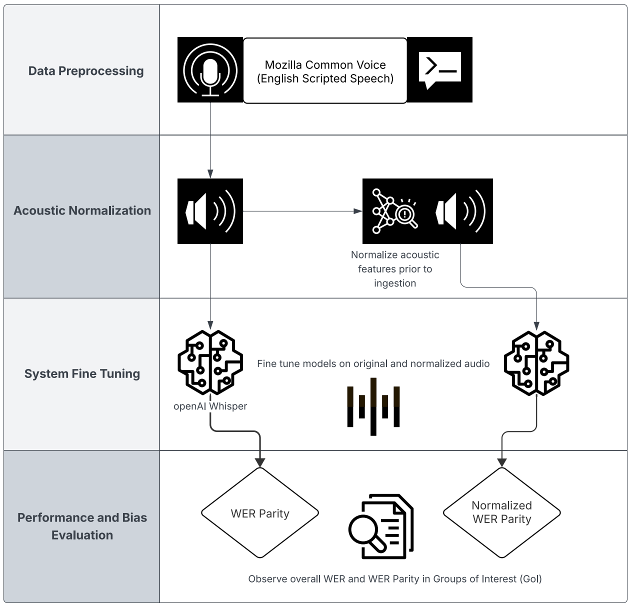
Data Preprocessing
Mozilla Common Voice (English Scripted Speech)
Acoustic Normalization
Normalize acoustic features prior to ingestion
System Fine Tuning
Fine tune models on original and normalized audio
openAI Whisper
Performance and Bias Evaluation
WER Parity
Normalized WER Parity
Observe overall WER and WER Parity in Groups of Interest (GoI)

Figure 1: GRAFN Pipeline

where μ is the estimated cepstral mean and Λ is the inverse diagonal covariance matrix, updated recursively to support real-time application [3].

Other methods exploit temporal adaptation to correct for dynamic speech properties such as speaking rate. Speech rate normalization adjusts the frame grid used to compute delta and delta-delta features so that phonetic transitions remain temporally aligned across different speaking speeds. This process is often applied in combination with speaker-specific spectral warping, as in vocal tract length normalization (VTLN), where the frequency axis is linearly or piecewise-linearly warped during feature extraction [4].

More advanced normalization frameworks explicitly model the transformation as an affine mapping in cepstral space. In acoustic space normalization,

$$\hat{x} = Lz + w, \tag{2}$$

a learned matrix L performs spectral normalization (e.g., speaker-dependent frequency scaling), while an additive vector w corrects environmental distortions [5]. Related work optimizes affine transforms using constrained maximum likelihood linear regression (CMLLR),

$$\hat{x} = Ax + b, \tag{3}$$

mapping each speaker into a normalized acoustic space to reduce speaker-induced variability during both training and inference [6].

Beyond strictly linear normalization, non-extensive statistical approaches propose alternative logarithmic transformations that better accommodate non-additive interactions between speech and noise. In particular, q-log spectral mean normalization replaces the conventional log operation with

$$\log_q(x) = \frac{x^{1-q} - 1}{1 - q}, \tag{4}$$

and performs mean normalization in this transformed domain, improving robustness by modeling cross-term effects explicitly [7].

Finally, real-time mean and variance normalization methods examine the trade-off between normalization accuracy and algorithmic delay, showing that dynamic estimation of statistical parameters can lead to performance gains but may degrade if insufficient look-ahead is permitted [8].

3 How it Relates to Bias in Gendered Voices

Bias mitigation strategies for systematic disparities in Automatic Speech Recognition (ASR) performance are commonly measured using the difference in word error rate (WER) between genders

$$\mathrm{WER}_{\mathrm{gap}} = \mathrm{WER}_{\mathrm{female}} - \mathrm{WER}_{\mathrm{male}}. \tag{5}$$

or statistical tests such as z-test and ks-test for such differences. These strategies often focus on direct balancing techniques at the sub-group level. Diversification is one such strategy, providing a balanced representation of types of speakers in the dataset [9]. This can be accomplished by oversampling underrepresented subgroups in the training dataset or under-sampling the number of training sames from over-represented subgroups [10]. Traditional augmentation techniques for oversampling speech data include time masking, background noise addition, and frequency masking. However, blanket increases or decreases the number of overall samples in speech data do not adequately balance data across demographic features. Bera and Agarwal implement two different versions of under-sampling: an overall method that targets an entire dataset with duration adjustments across sub-categories to ensure equal representation, and a duration interval-based method that performs balancing on divisions of the dataset separated by their duration [10].

Model-based approaches of measuring fairness investigate WER disparities across subgroups of interest by employing Poisson regression [11, 12]. Fairness assessments with respect to some factor of interest $f(\cdot)$, such as the gender/sex of the speaker, compute the effect of that factor on WER assessments across a number of different subgroups

$$C_s \overset{\text{i.i.d.}}{\sim} \mathrm{Poisson}(\lambda_s) \tag{6}$$

$$\log(\lambda_s) = \log(N_s) + \mu_{f(s)} \tag{7}$$

where C_s is the count of word errors (sum of insertion, deletion, and substitution errors), λ_s is the Poisson (mean) parameter, N_s is the number of words in the reference text for the sth utterance in the evaluation dataset, and $\mu_{f(s)}$ refers to the factor effect corresponding to the subgroup of $f(s)$ [12]. Mixed-effects Poisson regression, which deals in both fixed effect and random effects, treats the factor of concern (in this case, speaker gender) as a fixed effect. By treating other speaker affects (race, accent) as random, speakers contained in a dataset are in turn treated as randomly sampled from some larger collection, the characteristics of which are being estimated, to better measure and interpret WER differences among subgroups of interest.

Acoustic discrepancies: Prior work has demonstrated that statistically measurable acoustic differences between male and female speech can introduce disparities in ASR performance, motivating efforts to quantify and reduce such gender-conditioned variability at the signal or feature level. Recent analyses identify that female speakers exhibit consistently higher recognition accuracy across multiple open-source ASR models, with quantifiable differences in fundamental frequency (F_0) and high-frequency spectral energy concentration. The resulting gap highlights that spectral prominence above 2 kHz in female speech interacts differently with feature normalization pipelines, indicating that such discrepancies emerge before model ingestion [13].

Further evidence shows that multichannel speech enhancement models also process male and female speech differently at the phoneme level. Although utterance-level signal-to-interference ratio (SIR) differences may appear minimal, computing SIR independently per phoneme category reveals measurable performance gaps favoring female voices. This disparity is expressed as

$$\Delta\text{SIR} = \text{SIR}_{\text{female}} - \text{SIR}_{\text{male}}, \tag{8}$$

with consistently positive values across most vowel, plosive, and fricative classes, particularly under moderate noise conditions [14].

Complementary studies confirm that spectral normalization methods relying on global feature statistics may unevenly compensate male and female voices. Differences in pitch distribution and spectral tilt sensitivity cause male speech to be more susceptible to under-normalization in noisy environments, while female speech may exhibit over-normalization artifacts under high-frequency emphasis [15]. These observed disparities emphasize that normalization strategies designed without gender-conditioned statistical awareness can inadvertently reinforce or amplify existing recognition imbalances.

4 A New Robust Normalization Method for Data, Before Training, to Mitigate Gendered Voices

4.1 Problem Definition

Despite recent advances in Automatic Speech Recognition (ASR), persistent performance disparities between male and female voices have been repeatedly linked to statistical mismatches in raw acoustic distributions prior to model ingestion. These differences arise from naturally occurring variations in fundamental frequency (F_0), formant positioning, spectral tilt, and silence-to-speech energy ratios. Let $x(t)$ denote a raw speech waveform from a speaker belonging to gender group $g \in \{\text{male}, \text{female}\}$. The expected acoustic distribution is then expressed as $p(x \mid g)$, where empirical evidence confirms that $p(x \mid \text{male}) \neq p(x \mid \text{female})$ across multiple measurable acoustic dimensions. This mismatch propagates through feature extraction pipelines and yields biased downstream recognition performance.

4.2 Motivation from Prior Work

Previous normalization techniques have primarily focused on cepstral or spectral feature domains rather than directly addressing raw waveform discrepancies. Approaches such as mean-variance normalization, affine cepstral normalization, vocal-tract length normalization, and q-logarithmic transforms demonstrate that pre-model statistical conditioning can significantly reduce mismatch [3, 5, 7]. More recent gender-focused analyses confirm that measurable performance gaps emerge before model inference, driven by spectral imbalance and pitch-conditioned energy distribution [13, 14, 15]. These findings motivate a unified raw-waveform normalization strategy that adaptively equalizes gender-dependent acoustic distributions while preserving linguistic fidelity.

4.3 Dataset - Mozilla Common Voice

Mozilla Common Voice (MCV) is an open source crowd-sourced dataset consisting of scripted speech, spontaneous speech, and public domain prompts, sentences, and text from volunteers intended for research and development in ASR and language identification [16]. Common Voice's size, openness, demographic information, volunteer base, and emphasis on real-world speech patterns make it highly suitable for experimentation across one or more demographic subgroups. The full MCV Corpus as of June 24th, 2025, contains 33,816 recorded hours, of which 22,642 have been validated, in 137 languages. The 22.0 English version of the corpus contains 3,759 recorded hours of English speech from 98,938 voices in MP3 format, of which 2,724 hours are validated at the time of writing. Data collection and validation are community-driven: Volunteers record sentences from a given text prompt and provide optional metadata such as age, sex, accent, language variant, country of origin, or type of recording device used. Validation is performed when volunteers listen to the recording and vote either "Yes" or "No" to indicate whether the spoken words match the text in the given prompt. Exiting the validation pool requires at least 2 out of 3 votes.

4.4 Gender-Responsive Adaptive Feature Normalization (GRAFN)

Given an input waveform $x(t)$, GRAFN learns a transformation

$$\tilde{x}(t) = \mathcal{T}_\theta\big(x(t)\big) \tag{9}$$

such that the resulting distribution satisfies

$$p(\tilde{x} \mid \text{male}) \approx p(\tilde{x} \mid \text{female}), \tag{10}$$

while preserving all phonetic and intelligibility-relevant structure.

4.4.1 Step 1 — Acoustic Statistical Alignment

We estimate instantaneous statistics over short windows W centered at time t:

$$\mu_x(t) = \mathbb{E}_{\tau \in W}[x(\tau)], \quad \sigma_x^2(t) = \mathbb{E}_{\tau \in W}\big[(x(\tau) - \mu_x(t))^2\big]. \tag{11}$$

A locally normalized signal is produced via

$$x_{\text{norm}}(t) = \frac{x(t) - \mu_x(t)}{\sigma_x(t)}. \tag{12}$$

4.4.2 Step 2 — Adaptive Spectral Rebalancing

We apply an adaptive, learned spectral compensation to equalize gender-specific formant and energy distributions:

$$X_{\text{bal}}(f,t) = H(f,t) \cdot \mathcal{F}\{x_{\text{norm}}(t)\}, \tag{13}$$

where $\mathcal{F}\{\cdot\}$ is the short-time Fourier transform and $H(f,t)$ rebalances energy across critical bands known to diverge between genders.

4.4.3 Step 3 — Reconstruction and Output

The final normalized waveform is reconstructed as

$$\tilde{x}(t) = \mathcal{F}^{-1}\{X_{\text{bal}}(f,t)\}, \tag{14}$$

guaranteeing temporal alignment and waveform invertibility for downstream feature extraction.

GRAFN explicitly equalizes raw acoustic distributions across gender while preserving full linguistic fidelity, enabling model-agnostic deployment before any ASR feature pipeline.

4.5 Evaluation

We will evaluate GRAFN using gender-disaggregated word error rate (WER) and WER parity metrics under the framework of Raes et al. (2024)[17]. Performance will be measured across multiple architectures and noise conditions to validate whether pre-training waveform normalization can measurably reduce the gender recognition gap without degrading overall accuracy. Successful results would demonstrate that normalization, when executed at the raw acoustic level, can serve as a practical and principled pathway toward bias-resilient speech recognition.

5 Conclusion

This review establishes that gendered acoustic discrepancies in raw speech—particularly differences in pitch distribution, spectral tilt, and localized phoneme-level energy allocation—can propagate uncorrected into ASR pipelines and lead to systematic recognition bias. Existing normalization methods primarily operate at the feature or cepstral level and do not explicitly address pre-extraction statistical imbalances between male and female voices. Motivated by this limitation, we introduced *Gender-Responsive Adaptive Feature Normalization (GRAFN)*, a novel raw waveform–level normalization framework applied *before* any feature extraction or model training occurs.

GRAFN equalizes gender-conditioned acoustic distributions by combining three operations: (1) local statistical standardization to stabilize amplitude and energy variance, (2) adaptive spectral rebalancing to compensate formant and high-frequency asymmetries while preserving intelligibility, and (3) reconstruction into a fully invertible waveform suitable for any downstream ASR system. This design is explicitly model-agnostic and ensures no loss of phonetic fidelity during normalization.

6 References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," 2022. [Online]. Available: https://arxiv.org/abs/2212.04356

[2] A. Baevski, H. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," 2020. [Online]. Available: https://arxiv.org/abs/2006.11477

[3] O. Viikki, D. Bye, and K. Laurila, "A recursive feature vector normalization approach for robust speech recognition in noise," in *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP'98 (Cat. No. 98CH36181)*, vol. 2. IEEE, 1998, pp. 733–736.

[4] T. Pfau, R. Faltlhauser, and G. Ruske, "A combination of speaker normalization and speech rate normalization for automatic speech recognition," in *Proc. Int. Conf. on Spoken Language Processing ICSLP#, Beijing, China*, 2000.

[5] A. Acero and R. M. Stern, "Robust speech recognition by normalization of the acoustic space." in *icassp*, vol. 91, 1991, pp. 893–896.

[6] D. Giuliani, M. Gerosa, and F. Brugnara, "Improved automatic speech recognition through speaker normalization," *Computer Speech & Language*, vol. 20, no. 1, pp. 107–123, 2006.

[7] H. F. Pardede, K. Iwano, and K. Shinoda, "Feature normalization based on non-extensive statistics for speech recognition," *Speech Communication*, vol. 55, no. 5, pp. 587–599, 2013.

[8] P. Pujol, D. Macho, and C. Nadeu, "On real-time mean-and-variance normalization of speech recognition features," in *2006 IEEE international conference on acoustics speech and signal processing proceedings*, vol. 1. IEEE, 2006, pp. I–I.

[9] S. Feng, O. Kudina, B. M. Halpern, and O. Scharenborg, "Quantifying bias in automatic speech recognition," 2021. [Online]. Available: https://arxiv.org/abs/2103.15122

[10] A. Bera and A. Agarwal, "Bias detection and mitigation framework for asr system," in *7th International Conference on Signal Processing and Information Communications*, C.-C. Wang and R. G. B. Sangalang, Eds. Cham: Springer Nature Switzerland, 2025, pp. 13–27.

[11] M. Jahan, P. Mazumdar, T. Thebaud, M. Hasegawa-Johnson, J. Villalba, N. Dehak, and L. Moro-Velazquez, "Unveiling performance bias in asr systems: A study on gender, age, accent, and more," in *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2025, pp. 1–5.

[12] Z. Liu, I.-E. Veliche, and F. Peng, "Model-based approach for measuring the fairness in asr," 2021. [Online]. Available: https://arxiv.org/abs/2109.09061

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

[14] N.-E. Monir, P. Magron, and R. Serizel, "Evaluating multichannel speech enhancement algorithms at the phoneme scale across genders," *arXiv preprint arXiv:2506.18691*, 2025.

[15] S. Feng, B. M. Halpern, O. Kudina, and O. Scharenborg, "Towards inclusive automatic speech recognition," *Computer Speech & Language*, vol. 84, p. 101567, 2024.

[16] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," 2020. [Online]. Available: https://arxiv.org/abs/1912.06670

[17] R. Raes, S. Lensink, and M. Pechenizkiy, "Everyone deserves their voice to be heard: Analyzing predictive gender bias in asr models applied to dutch speech data," *arXiv preprint arXiv:2411.09431*, 2024.