Mitigating Gender Bias in ASR with Waveform–Level Feature Normalization

Dr. Cecilia Alm
Program Director Artificial Intelligence
Department of Psychology
coagla@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

***Abstract*—Automatic Speech Recognition (ASR) systems have advanced significantly over recent decades, yet their performance is still highly sensitive to acoustic variability. To combat gender bias in ASR, we propose 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] finetuned on Mozilla Common Voice, both before and after applying GRAFN to the training dataset.**

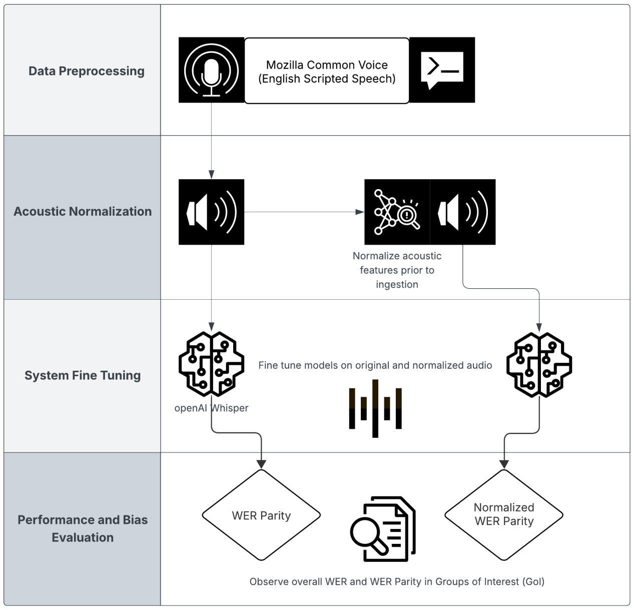
***Index Terms*—Word Error Rate, WER Parity, Cepstral Normalization**

I. Introduction

Automatic Speech Recognition performance possesses a sensitivity to variations in speaker characteristics. 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. Demographic features such as age, accent, and gender introduce further roadblocks to fair speech recognition when disproportionately underrepresented voices return higher rates of error. These biases pose harms in the form of reduced usability, inaccessibility, and are capable of reinforcing social inequalities when ASR technologies are deployed in real-world settings. 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. Using those works, it then proposes **Gender-Responsive Adaptive Feature Normalization (GRAFN)**, a model-agnostic approach aiming to reduce gender-based WER disparity while preserving phonetic and intelligibility-relevant structure.

A. 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 nonlinearities, address temporal variability such as speaking rate, or explicitly account for inter-speaker physiological differences through spectral warping or affine transformations [2]–[7]. 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.



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)

Fig. 1. GRAFN Pipeline

II. 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,

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

where μ is the estimated cepstral mean and Λ is the inverse diagonal covariance matrix, updated recursively to support real-time application [2].

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 [3].

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 [4]. 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 [5].

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 [6].

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 [7].

III. 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

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

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 [8]. This can be accomplished by oversampling underrepresented subgroups in the training dataset or under-sampling the number of training sames from over-represented subgroups [9]. 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 [9].

Model-based approaches of measuring fairness investigate WER disparities across subgroups of interest by employing Poisson regression [10], [11]. 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

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

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

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 $f(s)$ [11]. 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 [12].

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 [13].

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 [14]. These observed disparities emphasize that normalization strategies designed without gender-conditioned statistical awareness can inadvertently reinforce or amplify existing recognition imbalances.

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

A. 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.

B. 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 [2], [4], [6]. More recent gender-focused analyses confirm that measurable performance gaps emerge before model inference, driven by spectral imbalance and pitch-conditioned energy distribution [12]–[14]. These findings motivate a unified raw-waveform normalization strategy that adaptively equalizes gender-dependent acoustic distributions while preserving linguistic fidelity.

C. 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 [15]. 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.

D. 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.

1) Step 1 — Acoustic Statistical Alignment: We estimate instantaneous statistics over short windows W centered at time t:

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

A locally normalized signal is produced via

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

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.

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.

E. Evaluation

We will evaluate GRAFN using gender-disaggregated word error rate (WER) and WER parity metrics under the framework of Raes et al. (2024) [16]. 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.

V. Implementation

The architecture starts with the English portion of the Common Voice corpus, which provides paired audio clips and demographic metadata in a tab-separated file `validated.tsv`. The first processing stage loads this table with `pandas`, drops entries with missing `age`, `gender`, or `accents`, and attaches full file paths to the raw audio files stored in the `clips/` directory. At this point, we have a tabular representation where each row corresponds to one utterance with an audio path, transcript, and associated demographic attributes.

The English Scripted Speech MCV dataset contains over 2.5 million audio clips at the time of writing, of which 918,105 are annotated with gender, age band, and accent demographic information. Due to resource constraints, a subset of 2,000 waveforms and corresponding sentences are randomly selected given a seed. Of these, 200 clips are held out for validation and testing, with the remaining 1,800 used during training.

Before constructing the HuggingFace dataset, we run an *offline GRAFN training* phase. This step uses only the raw audio paths and gender labels to learn how to spectrally equalize male and female speech. We first normalize the corpus gender strings (e.g., `male_masculine`, `female_feminine`) into canonical labels “male” and “female”, and collect all utterances with a known binary gender. For each of these signals, GRAFN performs local time-domain normalization using short, overlapping windows to compute a sliding mean and variance, and then computes a short-time Fourier transform (STFT) to estimate the average magnitude spectrum per gender. These statistics are aggregated into a neutral “target” spectrum, from which we derive two gender-specific filters, $H_{\text{male}}(f)$ and $H_{\text{female}}(f)$, that map male and female spectra toward the same target distribution. This is a one-time, dataset-level calibration step that produces fixed filters stored inside the `GRAFNNormalizer`.

After this offline calibration, we convert the cleaned `pandas` DataFrame into a HuggingFace `Dataset` and cast the audio path column to the `Audio` feature type, which enables automatic loading and resampling at 16 kHz. We then split the dataset into train and test partitions, while preserving demographic columns such as `gender`, `age`, and `accents` for later analysis. A mapping function `prepare_dataset` is applied over each example in both splits and acts as the per-utterance front-end. For each batch row, we retrieve the waveform from the `Audio` object, normalize the gender label, and, if it is recognized as “male” or “female”, pass the waveform through the GRAFN transform using the pre-learned filter for that gender. This produces a bias-mitigated waveform $\tilde{x}(t)$, which is then fed into the Whisper feature extractor to produce log-Mel input features, while the reference transcript is tokenized into label IDs.

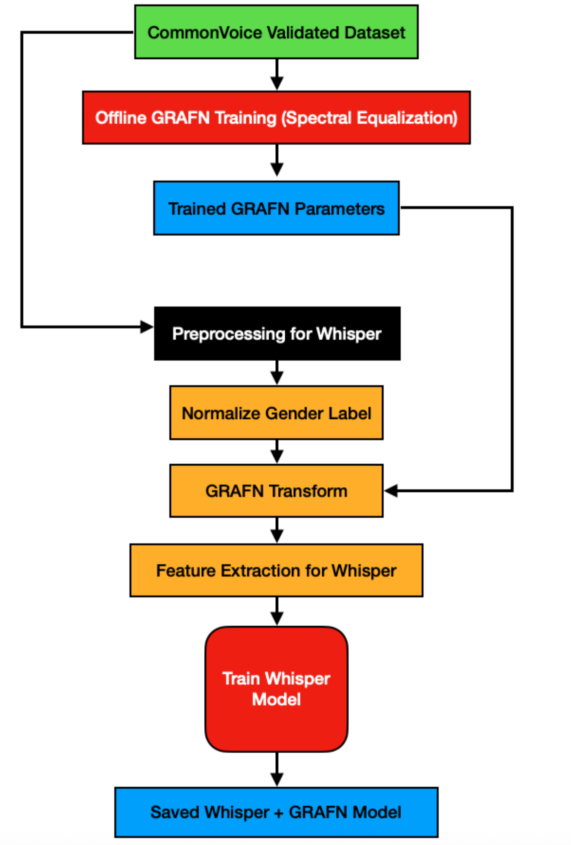

CommonVoice Validated Dataset
Offline GRAFN Training (Spectral Equalization)
Trained GRAFN Parameters
Preprocessing for Whisper
Normalize Gender Label
GRAFN Transform
Feature Extraction for Whisper
Train Whisper Model
Saved Whisper + GRAFN Model

Fig. 2. GRAFN-Whisper Architecture

These processed features drive a Whisper-tiny model equipped with low-rank adaptation (LoRA) modules. Concretely, we load the base `WhisperForConditionalGeneration` checkpoint, inject low-rank adapters into the attention projection layers (`q_proj` and `v_proj`), and train only these added parameters while freezing the remainder of the model. A custom data collator pads the variable-length input feature sequences and label sequences, replaces padding tokens in the labels with -100 so that they are ignored in the loss, and strips the beginning-of-sequence token from the targets where appropriate. The `Seq2SeqTrainer` is configured to optimize word error rate (WER) as the primary evaluation metric, using a `compute_metrics` hook together with `preprocess_logits_for_metrics` to decode predictions and references.

On top of the standard training loop, we attach a WER-parity callback to explicitly track fairness across demographic

Group	# Sentence	WER	GRAFN WER	WER Parity	GRAFN WER Parity
Male	1500	.1761	.2264	-.0044	.0207
Female	500	.1805	.2057		
Male	1000	.1784	.2347	.0082	.0093
Female	1000	.1702	.2254		
Male	500	.1753	.3372	-.0024	.1939
Female	1500	.1777	.1443		

TABLE I

WER AND WER PARITY OBSERVED IN DIFFERENT BALANCE CONFIGURATIONS OF RANDOMLY SELECTED SENTENCES FROM MOZILLA COMMON VOICE 23.0.

groups. At each evaluation step, this callback computes group-wise WER across the held-out test split, stratifying by the `gender` column. For each gender category present in the corpus (e.g., `male_masculine`, `female_feminine`), it filters the evaluation dataset, runs `model.generate` on the corresponding `input_features`, decodes hypotheses and references with the Whisper tokenizer, and computes a subgroup WER. It then reports both the per-gender WER values and a simple WER parity statistic (for example, the difference between male and female WER). This mechanism provides a direct lens on whether the GRAFN front-end and LoRA fine-tuning jointly narrow performance gaps between genders.

Finally, once training converges, the architecture includes a deployment stage. We first save the learned LoRA adapter weights, then reload the original base Whisper model and reconstruct a merged model by loading the adapters into the base checkpoint and calling `merge_and_unload`. The resulting merged model is moved to CPU and saved alongside the `WhisperProcessor` into a single deployment directory. At inference time, the same pipeline is reused: new utterances are passed through GRAFN (using the learned gender-specific filters), transformed into Whisper input features, decoded by the merged model, and, if desired, evaluated again with the same group-wise WER tooling. In this way, the implementation provides an end-to-end, model-agnostic gender-responsive normalization front-end (GRAFN) coupled with a bias-aware ASR training and evaluation pipeline.

VI. RESULTS

Initial WER parity refers to results obtained at the end of the fine-tuning process without applying GRAFN. In all observed scenarios where initial WER parity was negative, indicating a preexisting bias towards male speakers, this value became positive after applying GRAFN. However, the resulting WER gap's magnitude was often substantially higher, indicating that the resulting model now disproportionately biased against male speakers, and had become biased towards female. In all observed scenarios where initial WER was positive, indicating a preexisting bias towards female speakers, the WER gap after applying GRAFN would remain positive, with minimal increases in magnitude. GRAFN was found to adversely impact total WER after transform. Because initial experiments did not automatically balance the gender of speaker sentences selected, this was first believed to be caused by an imbalance in the training dataset, causing the model to overfit to female speaker characteristics. However, all imbalances in speaker count were overwhelmingly towards male (44% of annotated speakers self-report as male/masculine, and 18% of annotated speakers self-report as female/feminine) when randomly selecting sentences. Enforcing different levels of balance or imbalance demonstrate that GRAFN induces gender bias towards female speakers, even exacerbating WER disparity in cases of larger imbalance towards female speakers.

A. Limitations and Future Work

Under-utilization of Mozilla Common Voice: All experiments were carried out using a single Nvidia GeForce RTX 4070. Due to limited computational resources, only an extremely small (0.2%) subset of the English MCV 23.0 was used in the pipeline. Even after filtering for validated sentences with full entries for self reported age band, gender and accent, a method of additional restriction was required. A random seed for selecting sentence clips remains the optimal method to prevent user selection bias, but loses out on the diverse array of speaker characteristics present in the full dataset.

Additional Subgroups and Imbalances not Fully Explored: Initial experiments did not filter or balance for age, gender, or accent when selecting validated sentences. This was done to simulate natural conditions of a large and varied corpus of real world speech from a diverse collection of speakers. When GRAFN's effect on overall WER was first observed, ablation was subsequently performed by filtering for specific age bands and accents. This runs counter to the stated objective, which ideally included an analysis of multiple subgroups and their effect on WER bias.

Restricted to Whisper-Tiny: In addition to the time and computational limitations that impacted the size of Whisper chosen for the experiment, larger model sizes were observed to collapse during training. Typically after five epochs of LoRA adaptation using Whisper-small or Whisper-medium, WER and gradient normal would spike to extreme levels and remain so until final evaluation, often producing WER larger than 1.0. This trend was observed regardless of GRAFN application, indicating that the transform was not to blame. This initially led to the belief that the LoRA setup used in the experiment was not compatible with the larger Whisper models. Another likely explanation, though unverifiable, is a deprecation issue on the Windows operating system used to train the model with an older version of Pytorch (2.4), as it was the only version compatible with the latest available version of CUDA for the graphics processor installed on the machine.

VII. Ethics Statement

WER disparity between gender groups is a source of systematic bias in ASR performance. The objective of this paper - to mitigate WER disparity and improve fairness across demographic subgroups - is rooted in ethical considerations for groups disproportionately impacted by algorithmic bias, and remains its prevailing motivation. Additionally, use of the Mozilla Common Voice dataset requires an acknowledgment that users will not use speaker audio and demographic information to attempt to determine speaker identity. All analyses were performed at a group level: no attempt was made to identify individual speakers.

VIII. 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 seeks to equalize 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 intended to ensure no loss of phonetic fidelity during normalization. Evaluation of the GRAFN transform on randomly sampled subsets of the Mozilla Common Voice English dataset demonstrate a shifty in WER parity that favors female speakers. This trend persists across speaker gender imbalances, indicating that the transform preserves more semantic content in female speakers and maintains more speaker characteristics within distribution when applied to raw speech waveforms than male speakers.

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] 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.

[3] 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.

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

[5] 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.

[6] 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.

[7] 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.

[8] 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

[9] 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.

[10] 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.

[11] 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

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

[13] 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.

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

[15] 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

[16] 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.