"OK Aura, Be Fair With Me": Demographics-Agnostic Training for Bias Mitigation in Wake-up Word Detection

Fernando López[1,2], **Paula Delgado-Santos**[1]
Pablo Gómez[1], **David Solans**[1], **Jordi Luque**[1]
[1]Telefónica Innovación Digital, Madrid, Spain
[2]Universidad Autónoma de Madrid, Madrid, Spain
{fernando.lopez, paula.delgadodesantos, pablo.gomezguerrero,
david.solansnoguero, jordi.luque}@telefonica.com

Abstract

Voice-based interfaces are widely used; however, achieving fair Wake-up Word detection across diverse speaker populations remains a critical challenge due to persistent demographic biases. This study evaluates the effectiveness of demographics-agnostic training techniques in mitigating performance disparities among speakers of varying sex, age, and accent. We utilize the OK Aura database for our experiments, employing a training methodology that excludes demographic labels, which are reserved for evaluation purposes. We explore (i) data augmentation techniques to enhance model generalization and (ii) knowledge distillation of pre-trained foundational speech models. The experimental results indicate that these demographics-agnostic training techniques markedly reduce demographic bias, leading to a more equitable performance profile across different speaker groups. Specifically, one of the evaluated techniques achieves a Predictive Disparity reduction of 39.94% for sex, 83.65% for age, and 40.48% for accent when compared to the baseline. This study highlights the effectiveness of label-agnostic methodologies in fostering fairness in Wake-up Word detection.

Keywords: Wake-up word, fairness, bias, demographics-agnostic

1. Introduction

Voice-based interfaces are now central to human-computer interaction, enabling virtual assistants, hands-free messaging, and applications such as customer support and clinical/legal transcription. The entry point to most of these systems is a Wake-up Word (WuW); a predefined trigger phrase that, once detected by an always-on lightweight acoustic model, activates the device and initiates interaction with the user (Këpuska and Breitfeller, 2006; Këpuska, 2011; López et al., 2023; López-Espejo et al., 2021). However, speech systems often exhibit performance disparities across demographic groups such as sex, age, and accent (Attanasio et al., 2024). Acoustic variability can systematically affect and raise fairness concerns (Fuckner et al., 2023). In always-on settings, these disparities manifest not only as aggregate error-rate gaps but as unequal *interactional burdens* (Choi and Choi, 2025): some users must repeat commands or alter their speech more often than others to obtain the same functionality. WuW detection is particularly susceptible because decisions rely on short speech segments with limited context, which can amplify speaker-dependent variability and reduce reliability for children, older adults, and regional or non-native speakers.

These concerns are consistent with prior work documenting demographic bias across speech tasks, including speaker identification, phoneme recognition, intent classification, keyword spotting (KWS), and emotion recognition (ER) (Meng et al., 2022; Hutiri et al., 2023; Slaughter et al., 2023). In Automatic Speech Recognition (ASR), higher Word Error Rates (WER) are repeatedly reported for speakers with regional or non-native accents, with additional disparities linked to sex, age, and intersectional factors (Garg et al., 2018; Zolnoori et al., 2024; Harris et al., 2024; Feng et al., 2021; Martin and Wright, 2023). Evaluations of large foundation models such as Whisper (Radford et al., 2023) further corroborate persistent racial, sex, and dialect biases, frequently favoring majority or privileged groups (Fuckner et al., 2023; Slaughter et al., 2023; Hutiri et al., 2023). Similar patterns have been observed in KWS and ER, where systems often underperform for children, elderly speakers, and nonstandard accents (Mujtaba et al., 2024; Hutiri et al., 2023; Feng et al., 2021; Martin and Wright, 2023). Recent benchmark efforts such as Fair-Speech for ASR (Veliche et al., 2024) and FaiST for broader speech technology (Jahan et al., 2025) further document systematic performance gaps across multiple demographic attributes, underscoring the need for dedicated fairness analyses in speech interfaces.

Several methodological tools have been proposed to diagnose and mitigate these disparities. For instance, DivExplorer can automatically identify attribute combinations (e.g., sex, age, accent) associated with large performance gaps (Pastor

Proceedings of SPEAKABLE @ LREC 2026, pages 47–58
11 May 2026.

et al., 2021). Building on such diagnostics, mitigation frameworks such as CLUES use discovered subgroups to guide contrastive learning and reduce disparities by targeting underperforming cohorts in the representation space (Koudounas et al., 2024). In parallel, Slaughter et al. (2023) demonstrated that embeddings from pre-trained speech models, including Whisper, wav2vec 2.0, WavLM, and HuBERT, can encode and amplify social biases. To measure this directly within the embedding spaces, they developed the Speech Embedding Association Test (SpEAT). Building on this area of research, Lin et al. (2024) examined how specific architectural and data choices in self-supervised learning (SSL) impact social biases in downstream tasks.

Beyond specific methods and datasets, recent work argues that speech recognition fairness is inherently context-dependent and multi-metric rather than "one-size-fits-all", calling for benchmarks that consider task requirements, deployment constraints, and stakeholder needs (ElGhazaly et al., 2025; Veliche et al., 2024).

Nonetheless, many mitigation strategies depend on explicit demographic labels, which are often unavailable, incomplete, or privacy-sensitive in real-world deployments, and data scarcity for underrepresented groups remains a persistent obstacle (Dheram et al., 2022; Barocas and Selbst, 2016). Other approaches pursue fairness via personalization, for instance, by conditioning KWS models on speaker-specific embeddings to improve performance for underrepresented users (Labrador et al., 2025). While effective, such methods typically require additional user data and enrollment procedures, and may not be feasible for compact, always-on WuW detectors operating entirely on device.

Motivated by these limitations, we study demographic bias in WuW detection and develop mitigation strategies that do not require demographic labels during training. We (i) quantify demographic disparities across sex, age, and accent using group-wise analyses and fairness metrics such as Predictive Disparity (PD) and Disparate Impact (DI), and (ii) investigate demographics-agnostic training methods based on generalization-oriented data augmentation and knowledge distillation/transfer from large self-supervised foundation speech models. Experiments on a real-world Spanish WuW dataset ("OK Aura") show that these label-free strategies substantially reduce demographic bias while preserving overall detection performance.

2. Mitigation Methodology

We implement a mitigation pipeline that remains demographics-agnostic during training, reserving demographic labels strictly for post-hoc bias evaluation. First, we identify bias within the dataset to identify demographic groups underrepresented in the training and validation phases. Subsequently, we assess bias reflected in WuW classifier predictions. Then, we adopt demographics-agnostic training methodologies intended to alleviate those biases.

This choice is motivated by the high cost and practical barriers of collecting additional data for underrepresented groups (e.g., privacy and limited access). Even with balanced data, disparities can persist due to design choices (Hutiri et al., 2023) or feature selection (Bailey and Plumbley, 2021); while demographics-aware methods can reduce bias (Dheram et al., 2022), we target mitigation without explicit demographic conditioning.

Our methodology employs two demographics-unaware training strategies. First, we hypothesize that modulating or partially removing frequency information during training discourages the model from relying on demographic-correlated acoustic cues. Given that sex, age, and accent are known to correlate with F0 (fundamental frequency) and formant structure (Vorperian et al., 2019), spectral envelope (Harnsberger et al., 2008), and prosody (Piat et al., 2008), respectively. By disrupting these cues, the model could be encouraged to learn more invariant representations (Vandenberghe et al., 2023). To this end, we explore data augmentation techniques applied at the spectrum level (Section 2.1). Second, large pre-trained SSL models, trained on diverse audio, have been shown to suppress speaker identity in their upper layers (Mohamed et al., 2022). Furthermore, as some models have been scaled to encompass over 4 million hours of training data (Barrault et al., 2023), we hypothesize that they capture demographically robust representations. Therefore, we investigate using such models as teachers to train a compact, robust student model (Chai et al., 2022) (Section 2.2).

2.1. Data augmentation techniques

We consider both time-domain and time-frequency-domain augmentations. Given an input waveform $x \in \mathbb{R}^N$, we compute a time–frequency representation via the Short-Time Fourier Transform (STFT) and use its magnitude to form a spectrogram,

$$X = \mathrm{Spectrogram}(x) = |\mathrm{STFT}(x)|^2, \qquad (1)$$

where $X \in \mathbb{R}^{T \times F}$, T denotes the number of time frames, and F the number of frequency bins. For augmentations applied in the time-frequency domain, we modify the magnitude to obtain X' while preserving the original phase, and then reconstruct an augmented waveform x' using the inverse STFT (ISTFT).

FreqMixStyle: it mixes frequency-wise feature statistics between samples to promote domain-invariant representations. It is motivated by frequency-wise instance normalization analyses for audio domain generalization (Kim et al., 2022). In practice, we normalize a spectrogram X_i along the frequency axis and re-scale it using mixed statistics from another randomly selected spectrogram X_j:

$$\mu_{\text{new}} = \lambda\mu_i + (1-\lambda)\mu_j, \quad \sigma_{\text{new}} = \lambda\sigma_i + (1-\lambda)\sigma_j \tag{2}$$

where $\lambda \sim \text{Beta}(\alpha, \alpha)$ controls the interpolation strength.

FilterAugment: simulates acoustic filtering by applying smooth frequency-dependent gains rather than masking entire bands (Nam et al., 2022). Given a spectrogram X, we apply a multiplicative weighting mask

$$X' = X \odot W_{\text{FA}}, \tag{3}$$

where $W_{\text{FA}} \in \mathbb{R}^{T\times F}$ contains frequency-dependent weights and $\odot$ denotes element-wise multiplication. We use the linear variant, which linearly interpolates gains across frequency to avoid abrupt discontinuities (Nam et al., 2022).

Frequency masking: as a strong baseline augmentation, we also apply frequency masking from SpecAugment (Park et al., 2019). We sample the mask width $f \sim \mathcal{U}(0, W_F)$ and starting index $f_0 \sim \mathcal{U}(0, \nu - f)$, where ν is the number of mel channels, and set the band $[f_0, f_0 + f)$ to zero:

$$X' = \text{FreqMask}(X). \tag{4}$$

This technique has been shown to improve model robustness by forcing the network to learn from partial spectrograms, thus improving generalization. It is especially effective in situations where the model must handle varying acoustic conditions or incomplete audio inputs (Kim et al., 2021).

Device Impulse Responses (DIR). Impulse responses model how a capture device filters an input signal. Originally, it was presented for device generalization by simulating microphone characteristics. Nonetheless, it modulates frequencies by convolving each training utterance with a sampled device impulse response h_{dir} (Morocutti et al., 2023):

$$x' = x * h_{\text{dir}}. \tag{5}$$

To keep input dimensions consistent, we truncate the convolved signal to match the original length.

2.2. Speech Self-Supervised Learning models

SSL has become a key approach for learning robust speech representations from large amounts of unlabeled audio (Mohamed et al., 2022; Chen et al., 2022; Han et al., 2025; Wang et al., 2024). In WuW settings, SSL models can be particularly useful under limited labeled data and challenging acoustic conditions (Yu et al., 2023; Mørk et al., 2024).

We leverage a large SSL encoder to build a high-capacity teacher classifier (`w2v-BERT2-kws`) and then distill its knowledge into a compact WuW student model. Specifically, we use the w2v-BERT 2.0 pre-trained model (Barrault et al., 2023), a Conformer-based multilingual speech encoder trained on 4.5M hours of unlabeled audio. The `w2v-BERT2-kws` architecture is depicted in Figure 1. We leverage the w2v-BERT 2.0 encoder frozen and train a lightweight classification head. Motivated by evidence that different transformer layers encode complementary information (Pasad et al., 2023), we compute a learnable weighted sum over the 24 layerwise hidden states. The resulting sequence representation is then processed with Multi-Head Factorized Attention (MHFA), a parameter-efficient variant of multi-head attention that factorizes the attention projections so each head operates in a lower-dimensional subspace (Peng et al., 2025). Finally, the obtained result is summarized via attentive pooling before a final linear layer (Peng et al., 2025; Roncel Díaz et al., 2024). The teacher is trained with cross-entropy and used exclusively for distillation; it is not intended for real-time, on-device inference.

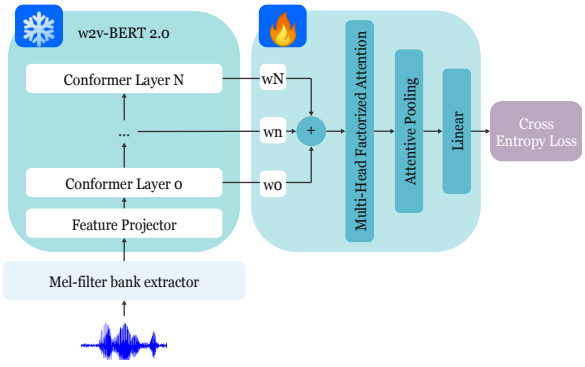

Figure 1: `w2v-BERT2-kws` architecture. Raw audio is converted to 80-channel Mel filterbanks, then passed through convolutional subsampling and a linear projection before a 24-layer Conformer encoder. Layerwise hidden states are combined via a learnable weighted sum, followed by Multi-Head Factorized Attention (MHFA), attentive pooling over time, and a linear classifier. The w2v-BERT 2.0 encoder is frozen; only the layer weights and the classification head are trained with cross-entropy.

Knowledge distillation (KD): after training the teacher, we freeze it and train the student via logit matching. The student minimizes a weighted combination of the standard cross-entropy (CE) loss

with respect to ground-truth labels and a Kullback–Leibler (KL) divergence term between temperature-scaled teacher and student predictions:

$$L_{\text{KD}} = \delta\, L_{\text{CE}}(p_{\text{student}}, y_{\text{true}}) + (1-\delta)\,\tau^2\, D_{\text{KL}}\big(p^{\tau}_{\text{teacher}} \parallel \log p^{\tau}_{\text{student}}\big) \tag{6}$$

where L_{KD} is the total loss function, and $\delta \in [0, 1]$ is a weighting factor that controls the balance between the two loss components. τ is the temperature parameter that controls the sharpness of the probability distribution, and $L_{\text{CE}}(p_{\text{student}}, y_{\text{true}})$ is the cross-entropy loss, defined as:

$$L_{\text{CE}}(p_{\text{student}}, y_{\text{true}}) = -\sum_i y_{\text{true},i} \log p_{\text{student},i} \tag{7}$$

where the term y_{true} represents the ground truth label and p_{student} is the probability output from the student WuW model. Here i refers to each specific class (WuW or unknown). The output probability is obtained by:

$$p_{\text{student},i} = \frac{e^{z_{\text{student},i}}}{\sum_j e^{z_{\text{student},j}}} \tag{8}$$

where $z_{\text{student},i}$ are the logits, and the j index refers to the summation over both classes. To continue with, $D_{\text{KL}}(p^{\tau}_{\text{teacher}} \parallel p^{\tau}_{\text{student}})$ is the KL divergence between two temperature-scaled probability distributions:

$$D_{\text{KL}}(p^{\tau}_{\text{teacher}} \parallel p^{\tau}_{\text{student}}) = \sum_i p^{\tau}_{\text{teacher},i} \log \frac{p^{\tau}_{\text{teacher},i}}{p^{\tau}_{\text{student},i}} \tag{9}$$

where $p^{\tau}_{\text{student},i}$ is the probability output from the student WuW model, obtained by applying a temperature-scaled softmax:

$$p^{\tau}_{\text{student},i} = \frac{e^{z_{\text{student},i}/\tau}}{\sum_j e^{z_{\text{student},j}/\tau}} \tag{10}$$

Similarly, p_{teacher} is also temperature-scaled:

$$p^{\tau}_{\text{teacher},i} = \frac{e^{z_{\text{teacher},i}/\tau}}{\sum_j e^{z_{\text{teacher},j}/\tau}} \tag{11}$$

where $z_{\text{teacher},i}$ are the logits from the teacher model and τ is the temperature parameter. Higher τ produces softer targets, encouraging the student to match relative class confidences rather than only hard decisions.

3. Datasets

We utilize a proprietary in-domain corpus, OK Aura (Section 3.1), and several publicly available out-of-domain resources for augmentation and robustness. Specifically, we incorporate Spanish Common Voice v7.1 (Mozilla Foundation, 2021), the M-AILabs Spanish corpus (Solak, 2019), real and simulated room impulse responses (RIRs) and noises from OpenSLR SLR28 (OpenSLR, 2016), and environmental noise recordings from DEMAND (Joachim Thiemann and Vincent, 2013). To further improve device robustness, we additionally use microphone impulse-response collections including MicIRP[1] and the Multi-Angle Multi-Distance Microphone IR dataset (Juan Carlos Franco Hernández, 2021).

Figure 2 summarizes how these resources are used across the experimental pipeline. OK Aura is used for training, validation, and testing. In contrast, the public corpora (Common Voice, M-AILabs, SLR28, DEMAND, MicIRP, and Multi-Angle Multi-Distance Microphone IR) are used primarily for training and validation to support augmentation and robustness. We restrict bias quantification and fairness evaluation to OK Aura because the out-of-domain datasets lack the necessary demographic metadata, provide insufficient granularity, or exhibit annotation mismatches relative to the WuW task.

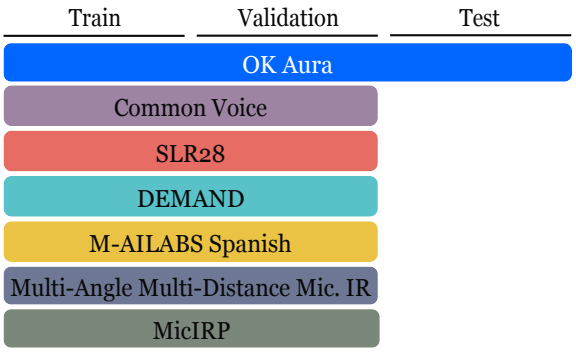

Figure 2: Dataset usage across train/validation/test splits. OK Aura is used in all phases (training, validation, and testing). Public resources are used primarily for training and validation, mainly for augmentation and robustness.

3.1. OK Aura Database

OK Aura contains approximately 5.8k audio samples (∼4.5 hours) from 546 anonymized speakers. It includes both speech and non-speech material (e.g., background noise), and all speech content is in Spanish, comprising the wake-up phrase and additional utterances. The corpus provides speaker-level demographic annotations (sex, age, accent) and sample-level metadata including class labels (positive/negative), transcriptions for speech samples, and a sound-type tag (“speech” vs. “noise”).

[1] https://micirp.blogspot.com/

To provide a concrete sense of the speech data, the corpus covers various realistic usage scenarios and challenging negative samples. Positive instances range from the isolated WuW (*"OK Aura"*) to the WuW embedded within context sentences (e.g., *"Perfecto, voy a mirar qué dan hoy. OK Aura"* 'Perfect, I am going to see what is on today. OK Aura'). Additionally, negative samples include utterances with partial matches, such as containing only the word "Aura" (*"Hay un aura de paz y tranquilidad."* 'There is an aura of peace and tranquility.') or "OK" (*"OK, a ver qué ponen en la tele."* 'OK, let's see what is on TV.'). Furthermore, it includes distractors with words that sound similar to the target phrase, such as *"Hola Laura,"* ('Hello Laura,'), *"Prefiero el hockey al baloncesto,"* ('I prefer hockey to basketball,'), or a combination of both: *"Porque Laura, ¿qué te pareció la película?"* ('Because Laura, what did you think of the movie?').

Furthermore, recordings span a wide range of acoustic environments as they were recorded in different spaces, from quiet rooms to natural background noise scenarios, and recorded across different devices. The dataset also includes temporal speech-event annotations (start/end times and total duration), obtained with the alignment procedure described by López and Luque (2022). A portion of OK Aura was released publicly as part of the Albayzin 2024 Wake-up Word Detection Challenge (Guillermo Cámbara and Segura, 2024) (López and Luque, 2024).

3.1.1. Demographic groups

For bias assessment, we consider three demographic attributes available in OK Aura: sex, age, and Spanish accent variety. We perform univariate analyses, evaluating each attribute independently. Sex is treated as a binary variable (Female/Male); age is grouped into 0–20, 21–30, 31–40, 41–50, and 51+; and accent labels cover the full annotation set: Unknown, Central Southern Spain, Southern Spain, Caribbean, Northern Spain, Northwestern Spain, Chilean, Eastern and Balearic Spain, Non-Native, Rio Plata, Canary Islands, Central America, Andean Pacific, Mexico, and Philippines.

3.1.2. Train and validation splits

We next analyze demographic distributions in the OK Aura training and validation splits to characterize representation imbalances. Table 1 reports the sex distribution, indicating a higher proportion of Male than Female samples. The age distribution has an average speaker age of 37 years, with most samples concentrated between 20 and 50 years old, and comparatively few samples from speakers under 20 or over 51 (Figure 3). Finally, accent labels are highly skewed, with Central Southern Spain dominating the training/validation data (Figure 4).

Sex	# Samples	Percentage
Female	2131	41.74%
Male	2974	58.26%

Table 1: Number of samples by Sex in the OK Aura Database (training and validation).

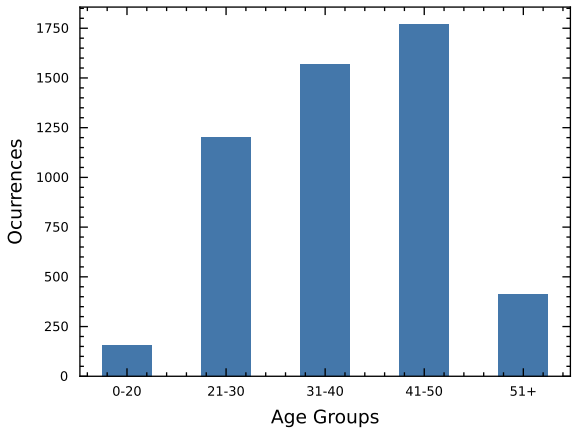

Figure 3: Age distribution in the OK Aura Database (training and validation).

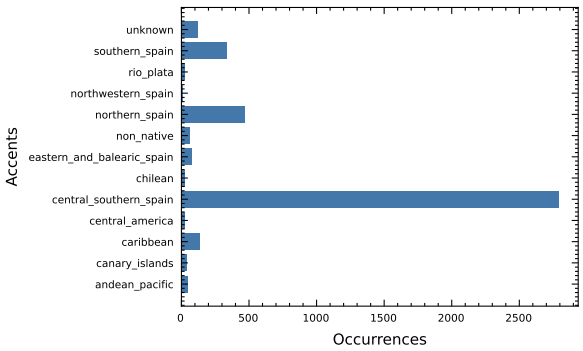

Figure 4: Accent distribution in the OK Aura Database (training and validation).

3.1.3. Test split

The OK Aura test split contains 575 samples of 47 unique speaker. Sex remains imbalanced (Table 2), and the age distribution is uneven, with a strong overrepresentation of middle-aged adults (41–50) and no samples from speakers aged 0–20 (Table 3). Accent coverage in the test set is also limited (Table 4), with several regional varieties either absent or severely underrepresented.

To ensure reliable subgroup estimates, we exclude demographic groups with fewer than 20 test

samples from bias quantification and fairness reporting. Furthermore, because some of the retained groups still feature a limited number of unique speakers (e.g., only 20 female speakers), subsequent fairness metrics should be interpreted as indicative trends rather than absolute guarantees of generalizability.

4. Experimental Setup

We first describe the WuW model, which is designed for on-device inference (Section 4.1) and the training procedure (Section 4.2). In the same section, we also detail how data augmentation and knowledge distillation are integrated into training. Finally, we define the metrics used to quantify data imbalance and predictive disparities (Section 4.3).

4.1. WuW detection model

We adopt a Gated Recurrent Unit (GRU)-based classifier as a practical trade-off between real-time efficiency and WuW detection accuracy. The model consists of a single GRU layer with 200 hidden units followed by a fully connected classification layer that discriminates between *WuW* and *unknown*. We refer to this architecture as `device-sgru`; it counts 145.6k parameters, and one inference takes $\sim$25 ms on a Pixel XL device (López et al., 2023), making it suitable for real-time inference on-device.

The input features are 13 MFCCs extracted with a 100 ms analysis window and a 50 ms hop, yielding 29 frames for a 1.5 s audio window. We replace the zeroth MFCC coefficient with the log-energy to better capture overall signal intensity (López et al., 2023).

For clarification, the `w2v-BERT2-kws` model described in Section 2.2 is just used to distill knowledge from it, it is not intended to be executed or deployed on device.

4.2. Training

All hyperparameters were set based on our previous research (López et al., 2023), prioritizing preservation of the strong baseline detection performance. Specifically, models are trained for up to 700 epochs by minimizing cross-entropy loss with a batch size of 128. We use Adam with an initial learning rate (LR) of 0.001 and reduce the LR by a factor of 10 when validation performance plateaus; training stops after four successive LR reductions without improvement.

To improve robustness under diverse acoustic conditions, we apply additive noise and reverberation (RIR convolution) during validation. Because such augmentation introduces additional variance in the validation loss, we select the final checkpoint as the one minimizing the mean of the three lowest validation-loss values across epochs, which stabilizes model selection under stochastic validation augmentation.

This procedure is used to train both the primary `device-sgru` model and the SSL-based teacher `w2v-BERT2-kws` model used for KD. The `device-sgru` model is trained from scratch with uniformly initialized weights; we refer to this configuration as `baseline`. For `w2v-BERT2-kws`, the w2v-BERT 2.0 pre-trained encoder is kept frozen, and only the task-specific layers are trained from scratch (uniform initialization).

We then integrate two demographics-unaware training strategies for bias mitigation: data augmentation and KD. During training, augmentations are applied with probability $p = 0.2$ (i.e., 20% of training samples), aiming to preserve strong baseline characteristics while injecting robustness-inducing perturbations.

The following augmentation configuration is used:

- **FilterAugment:** number of frequency bands uniformly sampled in $[3, 9]$, minimum bandwidth 187 Hz, gain sampled in ± 6 dB.
- **FreqMixStyle:** $\alpha = 0.4$ for the Beta distribution; mixing is restricted to pairs of samples with the same label.

Sex	# Samples	# Speakers
Female	254 (44.88%)	20
Male	321 (55.12%)	27

Table 2: Number of samples by Sex in the OK Aura Database (test).

Age Group	# Samples	# Speakers
0-20	0	0
21-30	135	11
31-40	138	11
41-50	295	24
51+	7	1

Table 3: Number of samples by Age Group in the OK Aura Database (test).

Accent	# Samples	# Speakers
central southern	313	26
eastern & balearic	15	1
non native	49	4
northern	90	7
southern	84	7
unknown	12	2

Table 4: Number of samples and speakers by regional Spanish accent in the OK Aura Database (test).

Attribute	Advantaged Group	Disadvantaged Group	Disparate Impact
Sex	Male	Female	0.7170
Age	41-50	21-30	0.6804
Accent	central_southern	northern	0.1692

Table 5: Disparate Impact by attribute in train and validation splits of OK Aura database.

- **Frequency masking:** $W_F = 30$ and $\nu = 128$ mel channels.

For KD, we initialize the student `device-sgru` with the weights of the pre-trained `baseline` to accelerate convergence. We then optimize the distillation objective in Eq. 6. During distillation we switch to Stochastic Gradient Descent (SGD) (momentum 0.9, weight decay 10^{-4}), as it can yield flatter minima and improved generalization. The initial LR is 0.0001 with an on-plateau scheduler, and we set $\delta = 0.2$ and $\tau = 2$.

4.3. Evaluation and metrics for bias quantification

Predictive disparities across demographic groups are often linked to data imbalance, where underrepresented cohorts tend to suffer degraded performance (Barocas and Selbst, 2016). We therefore relate demographic imbalance in the OK Aura training/validation data to disparities observed in model predictions on the test split. Concretely, we quantify imbalance in the input data (Section 4.3.1) and quantify predictive differences (Section 4.3.2).

4.3.1. Bias in data

To quantify demographic imbalance in the dataset, we use Disparate Impact (DI), a commonly used ratio-based metric in algorithmic fairness. Let G denote the set of demographic groups, with $a, d \in G$ representing an advantaged and disadvantaged group, and let $Y \in \{0, 1\}$ denote the binary label (1 for WuW presence and 0 otherwise). DI is defined as:

$$DI = \frac{P(Y = 1 \mid G = d)}{P(Y = 1 \mid G = a)}. \tag{12}$$

For multi-valued attributes (e.g., accent, age groups), we report the maximum ratio (or equivalently the most imbalanced pair) across all group pairs.

4.3.2. Bias in predictions

We follow the pairwise group comparison protocol described by Singh et al. (2023) to assess predictive disparities. We evaluate WuW detection on fixed 1.5 s windows and use a fixed decision threshold of 0.5. We report performance using the F1-score:

$$\mathrm{F1} = 2 \times \frac{\mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}, \tag{13}$$

where

$$\mathrm{Precision} = \frac{TP}{TP + FP} \tag{14}$$

$$\mathrm{Recall} = \frac{TP}{TP + FN} \tag{15}$$

Predictive Disparity (PD). We define PD as the maximum absolute difference in F1-score across demographic groups:

$$PD = \max_{i,j \in G} \left| \mathrm{F1}(g_i) - \mathrm{F1}(g_j) \right|, \tag{16}$$

where $g_i, g_j \in G$ denote group identities. Larger values indicate stronger performance gaps and potential fairness concerns.

Relative reduction of predictive disparity (RRPD). To compare mitigation strategies, we report the relative reduction in PD with respect to the baseline:

$$RRPD = 100 \times \frac{PD_{\text{baseline}} - PD_{\text{technique}}}{PD_{\text{baseline}}}. \tag{17}$$

Here, PD_{baseline} denotes disparity for the baseline model and $PD_{\text{technique}}$ for the model trained with a given mitigation technique.

5. Results and Discussion

We report (i) demographic imbalance in the OK Aura training/validation splits (data bias; Section 5.1.1), (ii) predictive disparities of the `baseline` WuW detector on the OK Aura test split (prediction bias; Section 5.1.2), and (iii) the impact of demographics-agnostic training strategies for mitigation (Section 5.2). Following our evaluation protocol, demographic groups with fewer than 20 test samples are excluded from bias quantification to ensure stable subgroup estimates.

5.1. Bias quantification

5.1.1. Bias in data

Table 5 reports DI for the OK Aura training/validation splits, revealing systematic representation imbalances across all examined attributes. Male speakers are overrepresented relative to female speakers ($DI = 0.717$), the 41–50 age group dominates compared to 21–30 ($DI = 0.6804$), and accent imbalance is most severe: Central Southern Spain is disproportionately represented relative

Group	F1-score	Support
Male	**0.9863**	296
Female	0.9825	204
21–30	**0.9956**	115
31–40	0.9828	118
41–50	0.9827	265
southern_spain	0.9818	84
central_southern_spain	**0.9873**	278
northern_spain	0.9781	70
non_native	0.9870	39
PD (sex)	0.0038	
PD (age)	0.0129	
PD (accent)	0.0092	

Table 6: Performance across demographic groups with predictive disparity (PD) for baseline model.

to Northern Spain ($DI = 0.1692$). These skewed distributions are likely to affect generalization and may translate into unequal predictive performance across groups.

5.1.2. Bias in predictions

Table 6 presents subgroup F1-scores and PD for the `baseline` model. Sex-related disparity is small but measurable ($PD = 0.0038$), with slightly higher F1 for male speakers (0.9863 vs. 0.9825). Age exhibits the largest performance gap ($PD = 0.0129$): the 21–30 group performs best (0.9956), while the 41–50 group performs worst (0.9827), highlighting older adults as a key cohort for mitigation. Accent disparities are also evident ($PD = 0.0092$), where Central Southern Spain achieves the highest F1 (0.9873) and Northern Spain is lower (0.9781), suggesting that accent variability remains a relevant source of error.

Overall, the baseline results indicate that demographic imbalance in the training data co-occurs with predictive disparities, motivating mitigation methods that increase robustness without requiring demographic labels.

5.2. Bias mitigation and analysis

First, the high-capacity SSL-based classifier `w2v-BERT2-kws` exhibits substantially lower disparities than `baseline` (Table 7), indicating that large-scale SSL pretraining can reduce, but not eliminate, demographic performance gaps. This motivates its use as a teacher model for KD.

Table 8 reports the relative reduction of predictive disparity achieved by each technique with respect to `baseline`. We observe that augmentation and KD show attribute-dependent behavior. Specifically, DIR only improves disparity for sex (67.35% RRPD), suggesting that device-specific impulse responses may not capture the heterogeneous acoustic variations typically associated with demographic attributes. FilterAugment yields the largest reduction in sex disparity (88.26% RRPD) and also improves age fairness (30.14% RRPD), but it increases accent disparity. This suggests that while smooth frequency-energy perturbations help reduce reliance on certain demographic-specific spectral cues, they do not universally benefit all attributes. On the other hand, FreqMixStyle improves age and accent fairness but degrades sex fairness (negative RRPD), indicating that frequency-wise statistics mixing affects demographic attributes differently and may not generalize across all cues simultaneously. We hypothesize that FreqMixStyle and FilterAugment, underperform on some speaker demographics as they can reshape frequency statistics too aggressively. They may destroy critical formant/prosodic cues and yielding mixed fairness gains at higher error cost.

In contrast, Frequency Masking provides the most consistent gains across all attributes, achieving a strong reduction in age disparity (83.65%) while also narrowing the gaps for sex and accent. Furthermore, Table 9 shows that these gains are achieved while maintaining competitive subgroup F1-scores, making Frequency Masking a suitable fairness-oriented augmentation. Suppressing specific frequency bands appears to force the model to distribute evidence across multiple regions of the spectrum rather than overfitting to a single demographic-correlated band (e.g., the F0 or low-formant region). This distributed attention both improves overall robustness and reduces reliance on demographic-specific cues.

Finally, KD reduces sex and age disparity but does not consistently reduce accent disparity. One plausible explanation is that accent invariance is constrained by limited accent diversity in the labeled in-domain data used during distillation, which may limit the teacher's ability to provide accent-neutral soft targets. Finally, combining KD with Frequency Masking does not improve over the best single-technique settings and can degrade results for some attributes, suggesting an interaction between stochastic spectral corruption and logit matching that may be difficult for a small student architecture to optimize jointly.

6. Conclusion and Future Work

This work shows that demographic-agnostic training can mitigate bias in Wake-up Word detection without requiring demographic labels during training. We studied two complementary families of methods: (i) data augmentation that perturbs or removes frequency information, and (ii) knowledge

Classifier	Sex RRPD (%)	Age RRPD (%)	Accent RRPD (%)
w2v-BERT2-kws	79.64	**85.35**	41.05

Table 7: w2v-BERT2-kws Relative Reduction of Predictive Disparity for demographic attributes in comparison to baseline model. It demonstrates reduced PD across sex, age, and accent categories.

Experimental Setting	Sex RRPD (%)	Age RRPD (%)	Accent RRPD (%)
DIR	67.35	0.00	-20.13
FreqMixStyle	-21.42	34.12	**40.48**
FilterAugment	**88.26**	30.14	-40.19
FreqMasking	39.94	**83.65**	**40.48**
KD	67.35	15.10	-20.13
KD + FreqMasking	21.24	15.10	-40.19

Table 8: Relative Reduction of Predictive Disparity (RRPD) across sex, age, and accent for different training techniques. Higher is better (negative values indicate increased disparity).

distillation (KD) from a large pre-trained speech model. Across speaker groups defined by sex, age, and accent, these approaches reduce performance disparities while maintaining competitive overall accuracy.

Our results highlight that augmentation design is critical for effective mitigation. In particular, frequency-energy perturbations and statistics mixing exhibited attribute-dependent behavior, sometimes degrading fairness for specific groups. Contrary, Frequency Masking emerged as the most robust single technique. It consistently achieved a relative reduction in predictive disparity (RRPD) of 39.94% (sex), 83.65% (age), and 40.48% (accent). By suppressing specific frequency bands, Frequency Masking prevents the model from overfitting to demographic-correlated acoustic cues (e.g., fundamental frequency) and forces it to distribute evidence across the broader spectrum. Additionally, KD achieved high RRPD, especially for sex and age, but showed limited impact on accent. This suggests that accent-invariant transfer remains constrained by the limited accent diversity available in the in-domain distillation data.

Future work will extend this analysis to intersectional fairness settings (e.g., older females with specific regional accents) and broaden demographic coverage by curating more balanced data across attributes. Particular emphasis will be placed on underrepresented age and accent groups.

Group	F1-score	Support
Male	0.9828	296
Female	**0.9851**	204
21–30	**0.9880**	115
31–40	0.9828	118
41–50	0.9847	265
southern_spain	0.9818	84
central_southern_spain	0.9835	278
northern_spain	0.9781	70
non_native	**0.9870**	39
PD (sex)	0.0023	
PD (age)	0.0052	
PD (accent)	0.0089	

Table 9: Predictive Disparity using FreqMasking technique to train device-sgru model.

Limitations. (i) Our analysis is univariate and does not capture intersectional effects (e.g., older female speakers with regional accents). (ii) The training/validation procedure includes out-of-domain audio sources, which can introduce distribution mismatch and metadata inconsistencies. (iii) Several demographic groups are underrepresented in the test split and were excluded from fairness reporting; furthermore, the limited number of speakers within the retained groups implies that our specific conclusions regarding them should be interpreted with caution. (iv) F1 does not decompose disparities into false accepts and false rejects, which carry asymmetric costs in WuW detection. (v) Conclusions at a fixed threshold of 0.5 may not generalize across operating points. Future work should adopt multi-objective evaluation frameworks that jointly optimize overall accuracy, per-group F1, and cross-attribute fairness, rather than treating each in isolation.

Ethics Statement This research addresses fairness in speech systems, which has positive ethical implications for reducing discrimination. The OK Aura dataset involves anonymized speakers with informed consent. Our demographics-agnostic approach specifically avoids requiring sensitive demographic labels during deployment, protecting user privacy. However, we acknowledge that bias mitigation techniques may have unintended effects on other demographic groups not examined in this study, and that univariate analysis may miss intersectional discrimination patterns.

7. Acknowledgments

This project has been partially funded by the Spanish Project 6G-RIEMANN (Grant Agreement No. 2022/0005420) and by the European Union's Horizon 2020 RIA ELOQUENCE project (Grant Agreement No. 101135916). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or European Commission-EU. Neither the European Union nor the granting authority can be held responsible for them.

8. Bibliographical References

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21318–21340, Miami, Florida, USA. Association for Computational Linguistics.

Andrew Bailey and Mark D Plumbley. 2021. Gender bias in depression detection using audio features. In *IEEE 29th European Signal Processing Conference (EUSIPCO)*, pages 596–600.

Solon Barocas and Andrew D. Selbst. 2016. Big data's disparate impact. *California Law Review*, 104(3):671–732.

Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Junyi Chai, Zhihao Wang, Jiaxin Chen, Hao He, Dawn Song, and Xia Li. 2022. Fairness without demographics through knowledge distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35.

Zhengyang Chen, Sanyuan Chen, Yu Wu, Yao Qian, Chengyi Wang, Shujie Liu, Yanmin Qian, and Michael Zeng. 2022. Large-scale self-supervised speech representation learning for automatic speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6147–6151.

Anna Seo Gyeong Choi and Hoon Choi. 2025. Fairness of automatic speech recognition: Looking through a philosophical lens. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):605–614.

Pranav Dheram, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. 2022. Toward fairness in speech recognition: Discovery and mitigation of performance disparities. *INTERSPEECH*.

Hend ElGhazaly, Bahman Mirheidari, Heidi Christensen, and Nafise Sadat Moosavi. 2025. Fairness in automatic speech recognition isn't a one-size-fits-all. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 19169–19178, Suzhou, China.

Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. 2021. Quantifying bias in automatic speech recognition. *arXiv preprint arXiv:2103.15122*.

Marcio Fuckner, Sophie Horsman, Pascal Wiggers, and Iskaj Janssen. 2023. Uncovering bias in asr systems: Evaluating wav2vec2 and whisper for dutch speakers. In *IEEE International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pages 146–151.

Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*, 115(16):E3635–E3644.

Jiangyu Han, Federico Landini, Johan Rohdin, Anna Silnova, Mireia Diez, and Lukáš Burget. 2025. Leveraging self-supervised learning for speaker diarization. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

James D Harnsberger, Rahul Shrivastav, William S Brown Jr, Howard Rothman, and Harry Hollien. 2008. Speaking rate and fundamental frequency as speech cues to perceived age. *Journal of voice*, 22(1):58–69.

Camille Harris, Chijioke Mgbahurike, Neha Kumar, and Diyi Yang. 2024. Modeling gender and dialect bias in automatic speech recognition. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 15166–15184.

Wiebke Hutiri, Aaron Yi Ding, Fahim Kawsar, and Akhil Mathur. 2023. Tiny, always-on, and fragile: Bias propagation through design choices in on-device machine learning workflows. *ACM Transactions on Software Engineering and Methodology*, 32(6):1–37.

Veton Këpuska and Jason Breitfeller. 2006. Wake-up-word speech recognition application for first responder communication enhancement. In *Sensors, and Command, Control, Communications,*

and Intelligence (C3I) Technologies for Homeland Security and Homeland Defense V, volume 6201, page 62011E. International Society for Optics and Photonics, SPIE.

Byeonggeun Kim, Seunghan Yang, Jangho Kim, Hyunsin Park, Juntae Lee, and Simyung Chang. 2022. Domain generalization with relaxed instance frequency-wise normalization for multi-device acoustic scene classification. *arXiv preprint arXiv:2206.12513*.

Gwantae Kim, David K Han, and Hanseok Ko. 2021. Specmix: A mixed sample data augmentation method for training withtime-frequency domain features. *arXiv preprint arXiv:2108.03020*.

Alkis Koudounas, Flavio Giobergia, Eliana Pastor, and Elena Baralis. 2024. A contrastive learning approach to mitigate bias in speech models. *arXiv preprint arXiv:2406.14686*.

Veton Këpuska. 2011. Wake-up-word speech recognition. In Ivo Ipsic, editor, *Speech Technologies*, chapter 12. IntechOpen, London.

Beltrán Labrador, Pai Zhu, Guanlong Zhao, Angelo Scorza Scarpati, Quan Wang, Alicia Lozano-Diez, and Ignacio Lopez-Moreno. 2025. Personalizing keyword spotting with speaker information. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yi-Cheng Lin, Tzu-Quan Lin, Hsi-Che Lin, Andy T Liu, and Hung-yi Lee. 2024. On the social bias of speech self-supervised models. In *Proceedings of INTERSPEECH*, pages 4638–4642.

Fernando López and Jordi Luque. 2024. Albayzin evaluation 2024: Wake-up word detection challenge.

Iván López-Espejo, Zheng-Hua Tan, John HL Hansen, and Jesper Jensen. 2021. Deep spoken keyword spotting: An overview. *IEEE Access*, 10:4169–4199.

Fernando López and Jordi Luque. 2022. Iterative pseudo-forced alignment by acoustic CTC loss for self-supervised ASR domain adaptation. In *Proceedings of IberSPEECH*, pages 46–50.

Fernando López, Jordi Luque, Carlos Segura, and Pablo Gómez. 2023. Robust wake-up word detection by two-stage multi-resolution ensembles. *arXiv preprint arXiv:2310.11379*.

Joshua L Martin and Kelly Elizabeth Wright. 2023. Bias in automatic speech recognition: The case of african american language. *Applied Linguistics*, 44(4):613–630.

Yen Meng, Yi-Hui Chou, Andy T. Liu, and Hung-yi Lee. 2022. Don't speak too fast: The impact of data bias on self-supervised speech models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3258–3262.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

Jacob Mørk, Holger Severin Bovbjerg, Gergely Kiss, and Zheng-Hua Tan. 2024. Noise-robust keyword spotting through self-supervised pretraining. *arXiv preprint arXiv:2403.18560*.

Tobias Morocutti, Florian Schmid, Khaled Koutini, and Gerhard Widmer. 2023. Device-robust acoustic scene classification via impulse response augmentation. In *IEEE 31st European Signal Processing Conference (EUSIPCO)*, pages 176–180.

Dena Mujtaba, Nihar Mahapatra, Megan Arney, J Yaruss, Hope Gerlach-Houck, Caryn Herring, and Jia Bin. 2024. Lost in transcription: Identifying and quantifying the accuracy biases of automatic speech recognition systems against disfluent speech. In *Proceedings of Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4795–4809, Mexico City, Mexico.

Hyeonuk Nam, Seong-Hu Kim, and Yong-Hwa Park. 2022. Filteraugment: An acoustic environmental data augmentation method. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

Daniel S Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D Cubuk, and Quoc V Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. *arXiv preprint arXiv:1904.08779*.

Ankita Pasad, Bowen Shi, and Karen Livescu. 2023. Comparative layer-wise analysis of self-supervised speech models. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Eliana Pastor, Luca De Alfaro, and Elena Baralis. 2021. Looking for trouble: Analyzing classifier behavior via pattern divergence. In *Proceedings of the International Conference on Management of Data*, pages 1400–1412.

Junyi Peng, Ladislav Mošner, Lin Zhang, Oldřich Plchot, Themos Stafylakis, Lukáš Burget, and Jan Černockỳ. 2025. Ca-mhfa: A context-aware multi-head factorized attentive pooling for ssl-based speaker verification. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Marina Piat, Dominique Fohr, and Irina Illina. 2008. Foreign accent identification based on prosodic parameters. In *INTERSPEECH*, pages 759–762.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning (ICML)*, volume 202, pages 28492–28518. PMLR.

Daniel Roncel Díaz, Federico Costa, and Javier Hernando. 2024. On the use of audio to improve dialogue policies. In *IberSPEECH*, pages 151–155.

Harvineet Singh, Fan Xia, Mi-Ok Kim, Romain Pirracchio, Rumi Chunara, and Jean Feng. 2023. A brief tutorial on sample size calculations for fairness audits. *arXiv preprint arXiv:2312.04745*.

Isaac Slaughter, Craig Greenberg, Reva Schwartz, and Aylin Caliskan. 2023. Pre-trained speech processing models contain human-like biases that propagate to speech emotion recognition. pages 8967–8989.

Loes Vandenberghe et al. 2023. Exploring data augmentation in bias mitigation against non-native-accented speech. *arXiv preprint arXiv:2312.15499*.

Houri K Vorperian, Raymond D Kent, Yen Lee, and Daniel M Bolt. 2019. Corner vowels in males and females ages 4 to 20 years: Fundamental and f1–f4 formant frequencies. *The Journal of the Acoustical Society of America*, 146(5):3255–3274.

Xin Wang, Héctor Delgado, Hemlata Tak, Jee-weon Jung, Hye-jin Shim, Massimiliano Todisco, Ivan Kukanov, Xuechen Liu, Md Sahidullah, Tomi Kinnunen, et al. 2024. Asvspoof 5: Crowdsourced speech data, deepfakes, and adversarial attacks at scale. *arXiv preprint arXiv:2408.08739*.

Mingdong Yu, Xiaofeng Jin, Bangxian Wan, and Guirong Wang. 2023. A few-shot speech keyword spotting method based on self-supervise learning. In *16th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–5.

Maryam Zolnoori, Sasha Vergez, Zidu Xu, Elyas Esmaeili, Ali Zolnour, Krystal Anne Briggs, Jihye Kim Scroggins, Seyed Farid Hosseini Ebrahimabad, James M Noble, Maxim Topaz, et al. 2024. Decoding disparities: evaluating automatic speech recognition system performance in transcribing black and white patient verbal communication with nurses in home healthcare. *JAMIA open*, 7(4):ooae130.

9. Language Resource References

Guillermo Cámbara, Jordi Luque, David Bonet, Fernando López, Mireia Farrús, Pablo Gómez and Carlos Segura. 2024. *Okey Aura Wake-up Word Dataset*. Zenodo, 1.1.0.

Maliha Jahan, Yinglun Sun, Priyam Mazumdar, Zsuzsanna Fagyal, Thomas Thebaud, Jesus Villalba, Mark Hasegawa-Johnson, Najim Dehak, and Laureano Moro Velazquez. 2025. Faist: A benchmark dataset for fairness in speech technology. In *Proceedings of Interspeech*, pages 1343–1347.

Joachim Thiemann, Nobutaka Ito and Emmanuel Vincent. 2013. *DEMAND: Diverse Environments Multi-Channel Acoustic Noise Database*. The Journal of the Acoustical Society of America.

Juan Carlos Franco Hernández, Tim Brookes, Enzo De Sena. 2021. *Multi-Angle, Multi-Distance Microphone Impulse Response Dataset*. Zenodo, 1.0.0.

Mozilla Foundation. 2021. *Common Voice Corpus (Spanish), version 7.1*. Mozilla Common Voice.

OpenSLR. 2016. *Room Impulse Response and Noise Database (SLR28)*. OpenSLR.

Imdat Solak. 2019. *The M-AILABS Speech Dataset*. M-AILABS.

Irina-Elena Veliche, Zhuangqun Huang, Vineeth Ayyat Kochaniyan, Fuchun Peng, Ozlem Kalinli, and Michael L Seltzer. 2024. Towards measuring fairness in speech recognition: Fair-speech dataset. *arXiv preprint arXiv:2408.12734*.