

Toward Fair Speech Technologies: A Comprehensive Survey of Bias and Fairness in Speech AI

Yi-Cheng Lin
Yun-Shao Tsai[†]
Kuan-Yu Chen[†]

Graduate Institute of Communication Engineering, National Taiwan University

f12942075@ntu.edu.tw
r14942093@ntu.edu.tw
f13942135@ntu.edu.tw

Hsiao-Ying Huang[†]

Graduate Institute of Electrical Engineering, National Taiwan University

monicahuangml1@gmail.com

Huang-Cheng Chou[†]

Shrikanth Narayanan

Ming Hsieh Department of Electrical and Computer Engineering, University of Southern California

huangchengchou@gmail.com
shri@usc.edu

Yu Tsao

Research Center for Information Technology Innovation, Academia Sinica

yu.tsao@citi.sinica.edu.tw

Jian-Jiun Ding

Graduate Institute of Communication Engineering, National Taiwan University

jjding@ntu.edu.tw

Hung-yi Lee

AI Center of Research Excellence, National Taiwan University

hungyilee@ntu.edu.tw

Abstract

Speech technologies are deployed in high-stakes settings, yet fairness concerns remain fragmented across tasks and disciplines. Existing surveys either adopt a general machine-learning perspective that overlooks speech-specific properties or focus on a single task, missing failure patterns shared across the speech domain. Synthesizing over 400 studies spanning generation and perception tasks and emerging speech-language models, this survey presents a unified framework that links formal fairness definitions to evaluation, diagnosis, and mitigation. We formalize seven fairness definitions adapted to the speech modality and organize the field’s conceptual expansion through three paradigms: Robustness, Representation, and Governance. We then ground evaluation metrics in the mathematical cores of these definitions, organizing them into six families and mapping each family back to the definitions it operationalizes. We diagnose bias sources along the speech processing pipeline, surfacing speech-specific mechanisms such as channel bias as a demographic proxy and annotation subjectivity in emotion labels. We systematize mitigation strategies across four intervention stages, mapping each to the diagnosed sources. Finally, we identify open challenges and propose directions for future research.

1 Introduction

Over the past few decades, speech processing technologies, encompassing Automatic Speech Recognition (ASR) (Yu & Deng, 2015; Li, 2022; Su et al., 2024; Lin et al., 2026c), Text-to-Speech (TTS) (Lin et al., 2026a; Huang et al., 2022a; Kaur & Singh, 2023), Speech Emotion Recognition (SER) (Madanian et al., 2023; Wu et al., 2024), Speaker Verification (SV) (Rosenberg, 1976; Zhang et al., 2022c), Speech Large Language Model (SLLM) (Cui et al., 2025; Arora et al., 2025; Huang et al., 2024; Yang et al., 2025), and so on, have transitioned from controlled laboratory environments to pervasive deployment in real-world settings.

Traditionally, progress in this field has been driven by engineering-centric metrics, such as minimizing Word Error Rate (WER) (Chen et al., 2025a), optimizing Real-Time Factor (RTF), or enhancing Mean Opinion Score (MOS) (Huang, 2025; Lin et al., 2025a; Ren et al., 2025; Ritter-Gutierrez et al., 2025). However, as these technologies increasingly permeate high-stakes domains such as healthcare (Larasati, 2025), criminal justice (Loakes, 2022), recruitment (Hickman et al., 2025), and personal assistance (Mahmood & Huang, 2024), purely technical performance indicators are no longer sufficient to capture the complex impacts these systems have on human society (Selbst et al., 2019a).

This inadequacy stems from a fundamental shift in the nature of deployment: speech technologies have evolved from supportive tools into institutional gatekeepers. When speech systems are deployed in high-stakes environments, they do not merely process data; they distribute resources and recognition (Naim et al., 2015; Cheng, 2011; Leix Palumbo & Prey, 2024). A failure in these contexts is not random noise but often a reproduction of existing societal inequities, privileging dominant acoustic profiles while systematically erasing or pathologizing marginalized voices. Beyond mere failure, models that enforce rigid, standard-language norms inadvertently demand acoustic assimilation, treating diverse cultural identities as deviations to be corrected (Choi & Choi, 2025). Therefore, fairness cannot be treated as a post-hoc optimization step but must be understood as a core reliability requirement (Cunningham et al., 2025; Zolnoori et al., 2024). Crucially, achieving this in the speech domain requires navigating physical constraints absent in other modalities; it demands addressing the unique properties of the speech signal that make fairness uniquely difficult to enforce (Benzeghiba et al., 2007).

1.1 Uniqueness of speech modality in fairness

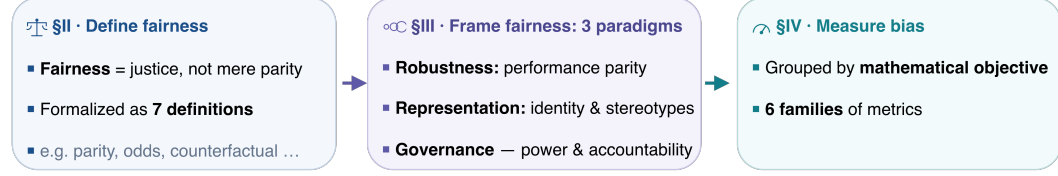
Speech presents distinct challenges for fairness because sensitive attributes are temporally and acoustically entangled with the linguistic content (Panariello et al., 2024; Raj et al., 2019). In text, linguistic units are discrete and editable, allowing sensitive terms to be removed at the token level (He et al., 2021; De-Arteaga et al., 2019; Sun et al., 2019). In images, sensitive cues such as faces or symbols may be partially localized, enabling region-level masking without necessarily destroying the entire visual context (Dehdashtian et al., 2024; Yu et al., 2021). Speech, by contrast, encodes lexical content, speaker identity, accent, affect, and other attributes jointly across time and frequency. Because these features are fused in the continuous signal, naive attempts to remove sensitive variables often degrade downstream utility or eliminate essential pragmatic cues, leaving no direct analogue to strategies such as removing a gender token or cropping a face.

Speech is also fundamentally temporal. Spoken interaction unfolds over time, so a system must decide who is speaking and when before it can interpret what was said (Castillo-López et al., 2025; Li et al., 2017). Diarization and endpointing rely on assumptions about turn structure and pause length that hold for typical adult conversation but not for everyone (Ryant et al., 2019): a person who pauses mid-utterance for cognitive reasons is cut off as if finished, and children or overlapping talkers are merged or dropped. Because these assumptions track particular populations, the resulting errors are not random but concentrate on the same groups. These failures concern who is heard at all, not just how accurately their words are transcribed, so a speaker can be disadvantaged even when the system would have recognized their speech correctly.

1.2 Difference from Prior Surveys

Existing surveys on fairness and bias fall into three broad categories: *general*, *Natural Language Processing (NLP)-focused*, and *speech-focused*. General fairness surveys (Mehrabi et al., 2022; Pessach & Shmueli, 2023; Oneto & Chiappa, 2020; Tang et al., 2023) provide foundational definitions and mitigation taxonomies but do not account for speech-specific properties such as acoustic entanglement or cascading errors across speech pipelines. NLP-focused fairness surveys (Blodgett et al., 2020; Gallegos et al., 2024; Bansal, 2022; Sun et al., 2019) address language-related bias yet treat text as the primary modality, overlooking the additional challenges introduced by continuous acoustic signals. Speech-focused reviews, in turn, often concentrate on a single task (e.g., ASR (Hinsvark et al., 2021; Ngueajio & Washington, 2022; Cunningham et al., 2025; Nakatumba-Nabende et al., 2026; Martin & Wright, 2023), health care (Larasati, 2025; Chand et al., 2023)) or a single risk dimension (e.g., privacy (Leschanowsky & Das, 2024)).

THE FRAMEWORK — how bias is defined (§II), how people frame it (§III), and we measure it (§IV)



ALONG THE SPEECH PIPELINE — where bias enters (§V) and is removed (§VI)

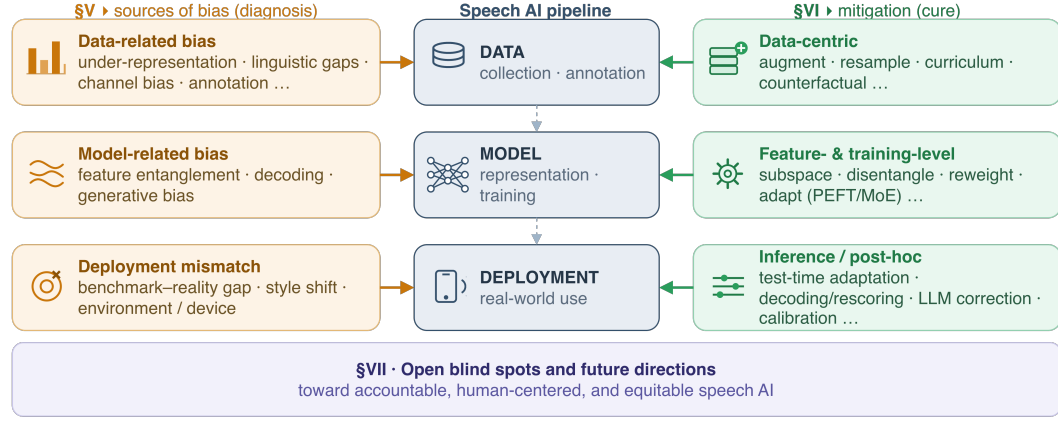


Figure 1: Overview of the survey, read top to bottom. The **upper half** is the framework: seven fairness definitions (§2), three paradigms (§3), and six metric families (§4). The **lower half** follows the speech processing pipeline, where each bias source (§5) sits directly above the mitigation that addresses it (§6); §7 lists the open problems.

Our survey differs from these in two respects. In *scope*, we jointly address speech production and perception tasks, as well as emerging speech-language models, enabling a cross-task perspective that reveals shared failure patterns invisible to single-task reviews. In *depth*, rather than descriptively cataloging known biases, we formalize seven definitions of fairness for the speech modality and build each subsequent layer of analysis on the preceding one: evaluation metrics are derived from the mathematical cores of these definitions, bias sources are diagnosed along the speech processing pipeline, and mitigation strategies are mapped back to the identified sources.

1.3 Main Contributions

To address the fragmentation in existing literature, this survey constructs a unified framework that connects formal definitions of fairness to computable metrics, bias diagnosis, and mitigation strategies across speech technologies. The four contributions below are organized as a coherent analytical pipeline: we first formalize *what fairness means* in the speech context, then specify *how to measure* it, *where bias originates*, and *how to mitigate* it, as summarized in Figure 1.

- **Formal Definitional Framework** (Sections 2 and 3): We formalize seven definitions of fairness adapted to the speech modality, spanning classical statistical notions and speech-specific desiderata. We further organize the conceptual expansion of speech fairness through three paradigms: Robustness, Representation, and Governance.
- **Grounded Evaluation Metrics** (Section 4): We organize evaluation metrics into six families defined by their shared mathematical objective, and map each family back to the fairness question it operationalizes, making explicit which metric answers which fairness question.
- **Speech-Specific Bias Diagnosis** (Section 5): We instantiate a pipeline-aware bias taxonomy, covering data, model, and deployment stages, with speech-specific mechanisms such as channel bias

as a demographic proxy, annotation subjectivity in emotion labels, and decoding assumptions that penalize atypical speech.

- **Mitigation Strategies** (Section 6): We systematize intervention techniques across four stages (data-centric, feature-level, training-level, and inference-stage), mapping each to the bias sources diagnosed above.

2 Background

Current academic and industrial trends indicate a *sociotechnical turn* in speech technology research (Selbst et al., 2019a). Scholars and practitioners are increasingly recognizing that technical systems do not operate in a vacuum (Blodgett et al., 2020); rather, they are deeply embedded within social structures (Sambasivan et al., 2021; Birhane, 2021). Consequently, these systems often exhibit significant performance disparities when interacting with diverse groups defined by race, gender, accent, or disability (Koenecke et al., 2020; Markl, 2022). To address these challenges, the academic community has begun to introduce and rigorously redefine two core concepts: Bias and Fairness.

2.1 Social Bias

While *bias* is a ubiquitous term in machine learning research, its definition remains frequently ambiguous and under-theorized. In traditional engineering contexts, the term *bias* typically refers to the error caused by simplifying a complex real-world problem into a model (Brynjarsdóttir & O’Hagan, 2014; Geman et al., 1992; Hastie et al., 2009). However, this technical definition is insufficient for capturing the ethical stakes of speech technology. Technical literature often fails to articulate exactly the harmful system behaviors, the populations they impact, and the normative ethical frameworks that justify these concerns (Blodgett et al., 2020). Without this normative grounding, *bias* risks being treated merely as a dataset imbalance to be corrected rather than a mechanism of social power that requires structural intervention. Therefore, this paper adopts the specific term **Social Bias** to describe systematic and unfair discrimination against individuals or groups based on socially salient characteristics such as race, gender, accent, or ability. By focusing on social bias, we move beyond technical errors to examine how speech technologies reflect and reinforce existing societal hierarchies.

2.1.1 The Mechanisms of Social Bias

Following the sociotechnical framework of Friedman and Nissenbaum (Friedman & Nissenbaum, 1996), we distinguish three mechanisms by which social bias enters a speech system (Lopez, 2021; Dobbe et al., 2018), ordered by when they act in the system’s life cycle. **Pre-existing Bias** is absorbed from society before any engineering begins. Its clearest instance in speech is the *Standard Language Ideology*, the societal norm that treats dominant varieties such as General American English as correct while stigmatizing others, such as African American English (AAE), as improper (Lippi-Green, 2012; Wolfram & Schilling, 2015); corpora collected under this norm teach models to treat non-standard speech as noise rather than valid communication (Tatman, 2017). **Technical Bias** arises during design, when seemingly neutral engineering choices carry discriminatory effects; a canonical example is telephony codecs, which are optimized for lower-pitched voices and degrade female speech before any model receives it (§5.3.3). **Emergent Bias** appears after deployment, when the usage context diverges from what designers assumed, as with a system trained on read speech encountering spontaneous or dysarthric speech (Tobin et al., 2024). Section 5 traces how these mechanisms materialize along the speech pipeline, from data collection through model design to deployment.

2.1.2 The Nature of Harm

To understand the impact of these mechanisms, it is necessary to distinguish between the types of damage they cause. Following Crawford and Barocas (Barocas et al., 2017), we distinguish two types of harm.

Table 1: Taxonomy of Allocative Harms in Speech Processing

Harm Type	Definition	System Manifestations in Speech Technology
Service Disparity	The system functions with significantly lower accuracy, reliability, or utility for specific social groups, imposing a functional tax or effective denial of service.	• High Error Rates: ASR systems exhibit significantly higher WERs for racial minorities (e.g., AAE speakers) and regional dialects compared to “Standard” speakers, rendering the technology unusable (Koenecke et al., 2020; Tatman, 2017). • Disability Exclusion: Voice interfaces such as voice assistants often fail catastrophically for speakers with dysarthria or stuttering, effectively locking people with speech disabilities out of the “voice-first” digital ecosystem (Sridhar & Wu, 2025; Lea et al., 2023; Mitra et al., 2021).
Opportunity Loss	The system serves as a barrier to formal opportunities, such as employment, housing, or financial services, by unfairly filtering candidates based on vocal characteristics.	• Hiring Discrimination: Automated video interview systems utilizing multimodal analysis infer competence from non-verbal cues. Adding audio features to these models decreases fairness, systematically penalizing candidates based on vocal prosody, while ASR errors on non-native speech introduce technical bias into the automated scoring pipeline (Booth et al., 2021; Hickman et al., 2025). • Educational Gatekeeping: English proficiency tests have incorporated automated speaking assessments. Test takers’ first-language backgrounds affect automated scores (Loukina et al., 2019; Kwako et al., 2023); such biases can make it more difficult for certain groups to meet thresholds required for admission or scholarships.
Economic Loss	The system’s failure or bias results in the deprivation of existing assets, direct financial costs, or reduced productivity (efficiency tax) for specific groups.	• Financial Lockout: Voice authentication systems rely on the assumption of vocal stability. Major physiological changes, such as those induced by aging, drastically alter acoustic features, triggering false rejections (Kelly et al., 2014; Kelly & Harte, 2011; Ai et al., 2025) that lock users out of their banking services. • Labor Displacement: Generative voice technologies (TTS/VC) decouple vocal identity from the physical speaker. The unauthorized extraction of professional voice actors’ data to train high-fidelity clones leads to direct labor substitution, as market-ready synthetic voices replace human labor in media production, devaluing the original speakers’ biological assets and rights to monetization (Sharma et al., 2025; Hutiri et al., 2024b; Berkowitz & Sweeney, 2025).
Physical Risk	The system’s failure to function in critical scenarios can result in physical harm, compromised health, or a delayed emergency response.	• Clinical Misdocumentation: ASR used for medical dictation exhibits higher error rates for non-native physician speech, leading to corrupted medical records and potential medication errors (Olatunji et al., 2023; Fatapour et al., 2025; Hodgson & Coiera, 2016). • Emergency Failure: Voice-activated emergency systems fail to recognize noisy, accented, or distressed voices during crises, delaying life-saving interventions (Galbraith et al., 2025; Aman et al., 2016; Moser et al., 2025).

Table 2: Taxonomy of Representational Harms in Speech Processing. This framework encompasses emerging harms associated with generative voice and recognition systems.

Harm Type	Definition	System Manifestations in Speech Technology
Stereotyping	The system reinforces essentialist, negative, or limiting generalizations about specific social groups through voice design or association.	• Digital Authority Bias: Voice assistants default to female voices for subservient roles, while "authoritative" or financial advice is synthesized in male voices, reinforcing gendered labor hierarchies (Ovacık, 2025). • Speaker Assignment Bias: synthesis models implicitly infer speaker gender from text prompts. When no speaker is specified, these systems tend to assign voices that align with occupational gender stereotypes (e.g., defaulting to female voices for "Nurse" prompts)(Puhach et al., 2025; Kuan & Lee, 2025).
Demeaning	The system generates content that mocks, degrades, or diminishes the dignity of specific groups (Denigration).	• Accent Caricature: TTS systems generating "accented" speech that relies on offensive stereotypes or "mimicry" rather than authentic linguistic patterns, creating a representation that users feel mocks their identity (Michel et al., 2025). • Algorithmic Stigmatization: ASR failures in processing AAE structures, such as invariant "be," lead users to attribute errors to the speaker's perceived lack of intelligence rather than system limitations, reinforcing deficit narratives of "bad grammar" instead of a legitimate linguistic variety (Martin & Wright, 2023).
Erasure	The system systematically fails to recognize or represent the existence of specific groups, rendering them invisible.	• Misgendering: Automatic Gender Recognition (AGR) systems that enforce a rigid male/female binary based on fundamental frequency, actively invalidating non-binary and transgender identities (Keyes, 2018). • Linguistic Erasure: ASR performance disparities that force speakers of non-standard dialects to "code-switch" to prestige dialects to be understood, delegitimizing their native speech variety (Markl, 2022).
Homogenization	The system collapses diverse sub-groups into a single, uniform representation, erasing intra-group cultural nuances.	• Accent Flattening: Data curation practices that label diverse regional varieties (e.g., Nigerian, Kenyan, Indian English) simply as "Non-Native," leading to models that treat these distinct acoustic cultures as a monolith (Markl, 2022; Prinos et al., 2024). • Global English Convergence: Multilingual and global ASR systems disproportionately optimize for a narrow set of dominant English varieties, treating linguistic diversity across World Englishes as noise rather than as structurally meaningful variation (Del Río et al., 2023).
Normative Bias	The system treats certain voices, speech styles, or interaction patterns as the default or normative standard, implicitly framing deviations as marked, deficient, or less appropriate.	• Unmarked Voice Standard: ASR and TTS systems implicitly center <i>neutral</i> or <i>standard</i> accents (often aligned with dominant social groups) as unaccented defaults, while labeling other speech varieties as <i>accented</i>, positioning them as deviations from an assumed norm (Bhattacharjee et al., 2025; Prinos et al., 2024). • Interactional Norm Bias: Voice interfaces are optimized for rapid, linear, and interruption-free speech, systematically disadvantaging speakers with neurodivergent communication styles, speech disfluencies, or culturally distinct turn-taking practices (Mujtaba et al., 2024; Raja et al., 2025).

Harm Type	Definition	System Manifestations in Speech Technology
Toxicity	The system generates content that is abusive, threatening, or hateful, or conversely, it systematically misidentifies neutral speech as having these characteristics due to bias against a specific identity.	• Identity-Based False Positives: Toxicity detection models disproportionately flag utterances containing terms related to marginalized identities (e.g., race, gender, religion) as hateful or abusive. This occurs even when the context is neutral or positive, resulting in the censorship of minority voices (Bell et al., 2025).
Cultural Appropriation	The unauthorized extraction and commodification of a marginalized group’s vocal identity for the benefit or entertainment of dominant groups, stripping the voice of its original cultural context.	• Digital Blackface: TTS or VC systems that enable non-Black users to adopt "Black" voices for social clout or entertainment, reducing a complex cultural identity to a performative costume without the lived experience of discrimination (Eidsheim, 2019). • Extractive Style Transfer: Generative models that clone the distinct vocal styles of indigenous speakers or specific artists without consent, treating their cultural expression merely as raw training data for commercial applications (Bird, 2020).

Allocative Harms. Allocative harms occur when a system unfairly distributes or withholds resources, opportunities, or information among different social groups. We categorize these harms based on the specific nature of the deprivation. A detailed taxonomy of these allocative failures specifically within the speech domain is summarized in Table 1.

Representational Harms. Representational harms arise when technical systems reinforce the subordination of specific social groups along the lines of identity. Unlike allocative harms, which concern the distribution of tangible resources or opportunities, representational harms affect the social standing of a group. These harms manifest when systems stereotype, demean, or erase a community, thereby influencing how that group is perceived, understood, and valued by society at large. Table 2 instantiates this taxonomy for speech, integrating harm categories from (Blodgett et al., 2020; Shelby et al., 2023; Weidinger et al., 2021; Suresh & Guttag, 2021; Gallegos et al., 2024).

2.2 Fairness

If *social bias* constitutes the diagnosis of a system’s pathology, ***Fairness*** represents the normative goal for its cure. In the machine learning literature, fairness is often reduced to a mathematical constraint (such as “equal performance across demographics”). Yet sociotechnical scholarship cautions that treating fairness purely as a technical specification can lead to abstraction traps, where designers narrow the problem in ways that erase the social context that gives the technology its meaning (Mulligan et al., 2019). Related work on language technologies further emphasizes that fairness claims must be explicit about ***what harm is being prevented, who is harmed, and which normative values justify the constraint***, because “fairness” is inherently value- and context-dependent rather than a single universal property (Gallegos et al., 2024).

Accordingly, this survey treats fairness for speech systems not as the mere absence of statistical bias but as a ***sociotechnical practice of justice*** (Selbst et al., 2019a). Because such a normative goal cannot be tested directly, it must be translated into a set of ***operational constraints*** (desiderata) stating which statistics, associations, or invariances a system should satisfy in a given application, and the selection of these constraints is itself a value-laden choice (Mehrabi et al., 2022).

Below, we formalize seven such constraints (Definitions 2–8), drawing on (Mehrabi et al., 2022; Selbst et al., 2019a; Dwork et al., 2012; Verma & Rubin, 2018); Definition 1 fixes the shared notation. They are ordered by what they constrain, moving from the system’s inputs, to its outputs, and finally to the task formulation itself.

Definition 1 (Speech System and Outcomes). *Let a speech system be a model M_θ that maps an input $X \in \mathcal{X}$ to an outcome $\hat{Y} = M_\theta(X) \in \mathcal{O}$ (such as transcript, score, class label). More generally, let $\mathcal{L}(\cdot)$ denote the law (distribution) of a random variable. When M_θ is stochastic, it induces an outcome distribution $\mathcal{L}(\hat{Y}) \in \Delta(\mathcal{O})$. Let $G \in \mathcal{G}$ denote a **social group** variable induced by one or more protected attributes. Let $Y \in \mathcal{O}$ denote the ground-truth outcome (reference transcript, true label, or target score).*

Definition 2 (Fairness Through Unawareness). *A speech system satisfies fairness through unawareness if protected attributes are not explicitly used in the decision process. Operationally, if $X = (X_{\text{audio}}, G)$ includes audio and explicit protected-attribute metadata G , unawareness requires that the system’s output does not change when G is removed:*

$$M_\theta(X_{\text{audio}}, G) = M_\theta(X_{\text{audio}}). \quad (1)$$

In speech, withholding G as an explicit input fails to prevent the model from relying on it, because accent, prosody, and voice quality often make G recoverable from the audio itself (Singh, 2019); the decision can still depend on G through these acoustic proxies, motivating the stronger constraints below.

Definition 3 (Demographic Parity). *Demographic parity requires that a system’s decisions are not affected by protected attributes. Let τ be a decision threshold chosen by the system operator, and define the binary decision $\hat{Z} = \mathbb{I}[\hat{Y} > \tau] \in \{0, 1\}$ (such as accept/reject in SV or a class decision in SER). Demographic parity requires:*

$$P(\hat{Z} = 1 \mid G = g) = P(\hat{Z} = 1 \mid G = g') \quad \forall g, g' \in \mathcal{G}. \quad (2)$$

More generally, parity can also be defined by distributional equality:

$$\mathcal{L}(\hat{Y} \mid G = g) = \mathcal{L}(\hat{Y} \mid G = g') \quad \forall g, g' \in \mathcal{G}, \quad (3)$$

or by bounded discrepancy under a distance D between distributions.

Definition 4 (Equalized Odds and Equal Opportunity). *Let $Z \in \{0, 1\}$ be the ground-truth binary label, and let $\hat{Z} = \mathbb{I}[\hat{Y} > \tau]$ be the predicted decision as in Definition 3. Equalized odds requires conditional independence of $\hat{Z}$ and G given Z :*

$$\begin{aligned} P(\hat{Z} = 1 \mid Z = z, G = g) &= P(\hat{Z} = 1 \mid Z = z, G = g') \\ \forall z \in \{0, 1\}, \forall g, g'. \end{aligned} \quad (4)$$

Equal opportunity is the special case for the positive class:

$$\begin{aligned} P(\hat{Z} = 1 \mid Z = 1, G = g) &= P(\hat{Z} = 1 \mid Z = 1, G = g') \\ \forall g, g'. \end{aligned} \quad (5)$$

These constraints are particularly relevant for thresholded biometric verification systems that trade off false accepts and false rejects.

Definitions 3–4 apply directly to binary or categorical classification tasks. For sequence-to-sequence tasks such as ASR, they require adaptation, for instance by defining the output as a thresholded WER indicator or applying the constraints at the token level. Sec. 4 provides the concrete metrics that operationalize these adapted constraints.

Definition 5 (Counterfactual / Invariance Fairness). *For any X with $G = g$, let $T_{g \rightarrow g'}(X)$ be the counterfactual input that preserves task-relevant content but transforms protected-attribute-linked characteristics to those of group g' . Let ψ be a dissimilarity function between outcomes (or outcome distributions). A speech system satisfies invariance up to ε if:*

$$\psi(M_\theta(X), M_\theta(T_{g \rightarrow g'}(X))) \leq \varepsilon \quad \forall g, g' \in \mathcal{G}, \quad (6)$$

where $\psi : \mathcal{O} \times \mathcal{O} \rightarrow [0, \infty)$ satisfies $\psi(a, b) = 0$ iff $a = b$. When M_θ returns distributions, ψ may be a distributional distance D :

$$D(M_\theta(\cdot \mid X), M_\theta(\cdot \mid T_{g \rightarrow g'}(X))) \leq \varepsilon'. \quad (7)$$

Unlike token swaps in text, the transformation $T_{g \rightarrow g'}$ must be physically realized in the waveform, typically by voice conversion or synthesis; current systems cannot yet alter one attribute in isolation, which limits how faithfully Def. 5 can be tested. *Definitions 2–5* assume access to clean, controlled inputs. In practice, speech systems operate under noisy, variable conditions. If these conditions correlate with demographics, robustness failures become fairness failures. *Definition 6* captures this insight.

Definition 6 (Robustness-as-Fairness Stress-Test Constraint). *Let ϕ_δ denote a realistic perturbation (task-irrelevant environmental/channel variations, noise) indexed by $\delta \sim \Pi$, where Π is a distribution over perturbation parameters. For an error/performance measure m , define the perturbed conditional risk:*

$$R_m^{\text{pert}}(g) = \mathbb{E}_{\delta \sim \Pi} \mathbb{E}[m(M_\theta(\phi_\delta(X)), Y) \mid G = g]. \quad (8)$$

Robustness-as-fairness requires bounded group disparities under perturbations:

$$|R_m^{\text{pert}}(g) - R_m^{\text{pert}}(g')| \leq \varepsilon \quad \forall g, g' \in \mathcal{G}, \quad (9)$$

or, as a complementary criterion, bounded disparity in degradation: given $R_m(g) = \mathbb{E}[m(M_\theta(X), Y) \mid G = g]$ and $\Delta_m(g) = R_m^{\text{pert}}(g) - R_m(g)$,

$$|\Delta_m(g) - \Delta_m(g')| \leq \varepsilon' \quad \forall g, g' \in \mathcal{G}. \quad (10)$$

Definition 7 (Output Association Equality). *When a speech system outputs text or feeds downstream language modules, let $s(\hat{Y})$ be a scalar association function (such as toxicity score, sentiment, gendered-term indicator). Association equality requires:*

$$\mathbb{E}[s(\hat{Y}) \mid G = g] = \mathbb{E}[s(\hat{Y}) \mid G = g'] \quad \forall g, g' \in \mathcal{G}, \quad (11)$$

or distributional equality

$$D(\mathcal{L}(s(\hat{Y}) \mid G = g), \mathcal{L}(s(\hat{Y}) \mid G = g')) \leq \varepsilon. \quad (12)$$

While *Definition 3* concerns the primary output itself, this definition targets secondary attributes derived from $\hat{Y}$ via an association function $s(\cdot)$.

Definition 8 (Validity and Contestability Constraint). *Let C denote the intended construct (“emotion”), and let L be the operational label/target. A minimal validity desideratum is measurement invariance across groups:*

$$P(L \mid C, G = g) = P(L \mid C, G = g') \quad \forall g, g' \in \mathcal{G}. \quad (13)$$

More broadly, the task specification should be contestable: stakeholders can interrogate whether L is an ethically and scientifically defensible operationalization of C even under high accuracy.

Because several of these constraints can conflict (e.g., demographic parity and equalized odds cannot hold simultaneously when base rates differ across groups (Barocas et al., 2023)), no single metric captures “fairness”; practitioners must select and justify a subset appropriate to their application context and stakeholder priorities. These definitions specify *what* properties to demand but leave open *along which dimensions* to evaluate them: at what granularity should groups be defined, and should the task formulation itself be scrutinized? The following subsection addresses these questions.

2.3 Fairness Dimension

We organize fairness concerns along two axes. **System Performance** asks whether outcome quality is uneven across speakers; **Problem Formulation and Validity** asks whether the task itself should be optimized at all.

2.3.1 System Performance Fairness

Here we characterize disparities in outcome quality as they surface at deployment. The literature examines such disparities at three granularities, termed **Group**, **Intersectional**, and **Individual** fairness. The three

differ chiefly in how finely the grouping variable G partitions the speakers, a parameter that Definitions 3–7 all rely on but leave unspecified. Broad demographic or linguistic categories yield group fairness, crossed attributes yield intersectional fairness, and a partition refined down to single speakers yields an individual-level test.

Group Fairness: Group fairness, the most widely studied granularity in the speech literature (Mehrabi et al., 2022; Chen et al., 2025b; Chien & Lee, 2024), instantiates G coarsely, partitioning speakers by demographic attributes (gender, age, race), linguistic attributes (accent, dialect, language), or contextual conditions (recording environment, device). Whichever axis is chosen, aggregate metrics such as overall WER or Equal Error Rate (EER) mask substantial inter-group gaps (Feng et al., 2021; 2024), motivating evaluation practices that report per-group performance and dispersion summaries before any mitigation is attempted (Patel et al., 2025; Liu et al., 2022; Tevissen et al., 2023; Lin et al., 2024b).

Given the diversity of these axes, a single system may simultaneously underserve several subpopulations to very different degrees. Aggregate metrics such as overall WER or EER can therefore mask substantial inter-group gaps: prior ASR audits show that disaggregating by speaker group reveals disparity patterns entirely invisible in corpus-level summaries (Feng et al., 2021; 2024). These observations motivate evaluation practices that explicitly report group-level performance and dispersion summaries, so that the *structure* of disparity is visible before any mitigation is discussed (Patel et al., 2025; Liu et al., 2022; Tevissen et al., 2023; Lin et al., 2024b).

Intersectional Fairness: Intersectional fairness examines combinations of attributes, recognizing that harms may concentrate in subgroups that are small, underrepresented, or acoustically distinct (Gohar & Cheng, 2023; Zee et al., 2024; Chen et al., 2025b). Intersectional gaps often arise when *multiple distribution shifts compound*, placing a subgroup at the margins of several groups that are underrepresented in or mismatched with the training distribution, where generalization is weakest. The resulting small sample sizes then make these concentrated errors harder to estimate and thus easier to overlook (Gohar & Cheng, 2023).

Recent multilingual ASR studies highlight this compounding: disparities can persist or even widen in worst-case slices despite improving average WER, and model scaling does not guarantee better fairness when the hardest intersectional groups remain under-covered or systematically mismatched (Zee et al., 2024). Similarly, cross-language audits show that the same demographic factor can manifest differently across languages and architectures; phoneme-level analyses link intersectional error patterns to articulation and phone confusions that are not captured by coarse group averages (Feng et al., 2024).

Individual Fairness: Beyond the per-speaker partition above, individual fairness in its general form requires that *similar* speakers or utterances receive similar outcomes, replacing the partition with a similarity measure (Dwork et al., 2012; Chen et al., 2025b; Mehrabi et al., 2022; Chien & Lee, 2024; Feng et al., 2023). In speech, violations arise when models latch onto nuisance variability such as speaking style, channel conditions, or speaker-specific expression (Verma & Rubin, 2018). In SV, two utterances from the same speaker may receive different scores primarily because their phoneme inventories differ (Baali et al., 2024); in SER, population-level models can perform well on average while consistently failing speakers whose expression departs from the training norm (Triantafyllopoulos & Schuller, 2024; Chou et al., 2024).

2.3.2 Problem Formulation and Validity Fairness

Definition 8 asks whether the task itself is a valid and ethically defensible thing to optimize, and its failure takes two forms in speech. A task may pursue a legitimate goal through an invalid proxy (**construct validity**), or it may target a construct that is contested in itself (**intrinsically unfair tasks**).

Construct Validity: Jacobs and Wallach (Jacobs & Wallach, 2021) formalize construct validity as the requirement that an operationalization faithfully captures the theoretical meaning of a construct; when it does not, “improving accuracy” can entrench an invalid measurement. A concrete example in speech is automated interview scoring, where “hirability” prediction models use prosodic cues such as pauses and intonation, sometimes explicitly recommending that candidates “speak more fluently” (Rasipuram et al., 2016; Naim et al., 2018). Yet fluency is not a neutral skill: employment contexts already penalize disfluent speakers, so optimizing for fluency risks laundering an existing discriminatory norm into an algorithmic target rather

than measuring genuine job competence (Klein & Hood, 2004). The result is a validity-driven fairness failure: a model can achieve group error parity while the metric it optimizes is itself an unfair proxy (Mujtaba et al., 2024).

Intrinsically Unfair Tasks: Some speech tasks are contested not because the proxy is poor but because the construct itself is ethically problematic. *Automatic Gender Recognition* (AGR) from voice reduces gender to a binary physiological inference, which systematically conflicts with transgender and non-binary identities and can impose representational harm through misgendering, even when a system reports strong aggregate accuracy (Keyes, 2018; Hamidi et al., 2018). Here, improving performance does not resolve the fairness issue: the harm stems from inferring and enforcing an identity label that many people do not consent to, a problem rooted in the task definition rather than in model quality.

The concepts introduced in this section provide the formal vocabulary for the remainder of this survey. However, the speech fairness literature did not adopt these concepts simultaneously; different research communities have foregrounded different diagnoses of what makes a system unfair. The next section traces three such paradigms and how they coexist in current research.

3 Expansion of Fairness in Speech

Based on a thematic review of the speech fairness literature, we identify three coexisting paradigms, distinguished by what each treats as the primary source of unfairness. **Robustness** attributes unfairness to distributional mismatch, seeking to close error gaps across speaker demographics, recording conditions, and deployment environments through broader data coverage and model adaptation. **Representation** attributes unfairness to the encoding of social meaning, asking whether learned representations and model outputs perpetuate stereotypes or erase marginalized identities. **Governance** attributes unfairness to power asymmetries in the sociotechnical context, addressing who collects the data, who defines the task, and who is accountable for downstream harms.

3.1 Fairness as Robustness: Bridging Performance Gaps in Recognition

In the initial paradigm identified through our analysis (emerging prominently from the mid-2010s, and still widely practiced today), fairness was most commonly operationalized as **performance parity** (Tatman, 2017). Research focused on auditing discriminative tasks such as ASR and SV, quantifying fairness through disparities in standard error metrics, including WER and EER (Shen et al., 2022). Seminal work highlighted performance gaps for AAE speakers (Tatman, 2017; Koenecke et al., 2020), and similar challenges were documented across other languages and dialects (Yang et al., 2022b). These disparities were largely framed as **robustness failures**: demographic differences were treated as distributional shifts that impaired generalization (Hinsvark et al., 2021), and the primary remedy was data diversification to narrow inter-group error gaps.

At a deeper conceptual level, however, this robustness-oriented framing carried a strong implicit assumption: inter-group variation was treated as noise to be reduced rather than as a meaningful expression of speaker identity. While this body of work was instrumental in establishing baseline benchmarks and making algorithmic bias empirically visible, it operationalized fairness primarily as accuracy-based parity. This framing centers on measurable error gaps rather than the social meaning or structural conditions of unequal outcomes. As a result, broader questions concerning deployment contexts, social meaning, and representational harm remained largely outside the scope of analysis, even in cases where aggregate error rates appeared to be equalized. The paradigms that follow take up these broader questions, widening what fairness is taken to include rather than correcting robustness.

3.2 Fairness as Representation: Mitigating Essentialism in Profiling and Generation

As speech technologies expanded beyond transcription into tasks that infer or generate speaker characteristics, a different category of fairness risk emerged: rather than merely failing to serve certain speakers, systems may actively mischaracterize or impose speaker identities. Accordingly, the focus of fairness research expanded beyond accuracy-centric evaluation to address essentialism, profiling, and representational harm.

Profiling and anti-essentialism: The rise of paralinguistic tasks, including SER and speech-based clinical assessment, introduces new risks associated with *profiling*: systems infer latent personal attributes (such as emotional state, health conditions, or employability) from observable vocal characteristics (Singh, 2019). These inferences rest on an *essentialist* assumption, namely that the mapping from acoustic features to personal attributes is fixed and universal across speakers (Hanna et al., 2020). In practice, however, vocal expression is culturally situated and individually variable (van Rijn & Larrouy-Maestri, 2023), so a mapping learned from one population does not generalize as a neutral ground truth. Evidence of systemic bias spans modalities: automated affective analysis has been found to misinterpret Black individuals as angrier (Rhue, 2018), while ASR performance degrades because training corpora underrepresent distinctive features of African American Language, such as habitual ‘be’ (Martin, 2021). The problem is not merely lower accuracy for AAE speakers, but that the model encodes a culturally specific norm as if it were an objective link between voice and emotion. In high-stakes contexts such as automated hiring and medical decision support, such essentialist mappings can lead to allocational harms through the systematic misinterpretation of socially meaningful vocal variation (Bouazizi et al., 2023).

Representation and stereotyping: Concurrently, the increasing ubiquity of TTS and VC technologies prompted the speech research community to initiate a critical discourse regarding representation in generated speech (Hutiri et al., 2024a). Unlike discriminative models, generative models actively shape the voices and interaction styles presented to users, raising concerns about representational harms. Fairness considerations in this context, therefore, extend to whether synthesized voices reproduce or amplify existing social hierarchies and stereotypes. For example, early deployments of voice assistants were frequently (Strengers & Kennedy, 2020; Phan, 2019) designed with compliant female-coded vocal personas, reinforcing gendered assumptions about servitude and care work (West et al., 2019). Similarly, VC and accent-altering systems have raised concerns about identity erasure, as the vocal distinctiveness of minority speakers can be normalized to standardized target voices, resulting in suppression of socially and culturally significant variation (Michel et al., 2025; Payne et al., 2024).

Profiling and generative stereotyping are two directions of the same representational failure. In both, the system treats the link between vocal signal and social identity as a fixed natural fact rather than something socially constructed. Profiling does this on the way in, reading identity from a voice, while generation does it on the way out, writing identity into a voice.

3.3 Fairness as Governance: Navigating Systemic Risks in Sociotechnical Pipelines

Current research on fairness in speech technologies has shifted from focusing on isolated model components to examining the *sociotechnical ecosystems* in which these technologies are deployed. Because technical architectures and design choices also shape who holds agency over a system’s behavior (Lessig, 1999; Winner, 1980), we treat *governance* broadly, beyond institutional regulation alone. We organize the resulting fairness risks across three interrelated scales: the pipeline scale of accountability, the interaction scale of human–AI power dynamics, and the institutional scale of data governance.

Pipeline scale: Composite systems and the accountability gap. In real-world deployments, speech technologies rarely operate as standalone modules. They are assembled into multi-stage pipelines that link recognition, language processing, and synthesis, and these components are frequently built and owned by different parties. This assembly, not any single model, is where a distinct fairness concern arises. When a harm emerges from the interaction of many components, responsibility for it is spread across *many hands* (Nissenbaum, 1997), so no single party is answerable and the affected speaker has no clear place to seek a correction or hold anyone responsible (Cooper et al., 2022). This is a part of fairness that no amount of technical accuracy can resolve (Selbst et al., 2019b), and it is what our contestability constraint (Definition 8) names. Cascading errors make the harm concrete. A high word error rate in an upstream transcription component can constrain a downstream language model’s interpretation of spoken intent (Ezema et al., 2025), leaving the speaker unable to make themselves understood, yet the resulting failure belongs to no component in isolation. Improving the accuracy of any one stage does not settle who is accountable when the assembled system fails a speaker.

Interaction scale: Digital authority and automation bias. Beyond system architecture, recent work also highlights the *power dynamics* embedded in human–AI interaction. Whereas previous research emphasized stereotyping in model outputs, current analyses focus on the *digital authority* of speech-based systems, understood as users’ tendency to defer to automated voice interfaces due to their perceived neutrality or expertise. Scholars argue that interaction design choices, including the use of standardized accents, authoritative speaking styles, or gendered personas, can function as sociotechnical control mechanisms that discourage questioning or critical engagement (Pradhan et al., 2025), thereby eroding user agency. In high-stakes settings, such dynamics contribute to automation bias (Cummings, 2017), whereby system outputs are treated as directives rather than fallible recommendations, reinforcing existing social hierarchies.

Institutional scale: From extraction to sovereignty. At the institutional level, early efforts to improve coverage focused on collecting speech from underrepresented accents and speaker groups. Such collection, however, was sometimes carried out without community consent or reciprocal benefit, turning data diversification into an *extractive practice* (Bird, 2020). More recent scholarship responds by emphasizing *data sovereignty*: marginalized and Indigenous communities have a right to determine how their linguistic data are collected, represented, and used (Papakyriakopoulos & Xiang, 2023; Leoni et al., 2024; Rajab et al., 2025). Fairness is thus reframed not as a problem of maximizing data diversity alone, but as a question of reclaiming community agency through institutional legitimacy, consent, and participation.

This institutional tension is one instance of a broader pattern. The three paradigms are not mutually exclusive, and a single practice can sit differently under each. Data diversification is the clearest example. What the robustness paradigm (§3.1) treats as the remedy for demographic error gaps, the governance paradigm sees as an extractive practice when it is done without community consent. A system is therefore best examined through all three lenses at once, since an intervention that answers one paradigm can raise a problem under another.

4 Evaluation Metrics

To operationalize the fairness paradigms of § 3, concrete evaluation metrics are needed. Because metrics that appear unrelated often share the same mathematical structure, while similar formulas can serve different normative goals, we organize metrics by their *mathematical objective* rather than by paradigm or task, yielding six *mathematical families* (Table 4). Five families group metrics around a shared formal quantity; the sixth (§ 4.6) collects metrics whose forms overlap with earlier families but whose evaluation target, the content of generative output, warrants separate treatment. Each family is annotated with its primary paradigm affiliation (Robustness, Representation, or Governance), making explicit where the same formula serves different normative purposes.

Table 3 maps these six families back to the fairness desiderata introduced in §2, showing which definition each family directly operationalizes (●) or partially reflects (◐). Definitions 2 and 8 are omitted because they operate at a meta-level. Definition 2 (Unawareness) requires only that G not be used explicitly, but in speech, acoustic features are strong proxies for demographics, so every family in this section implicitly goes beyond that baseline. Definition 8 (Validity) questions whether the task formulation and label taxonomy are themselves defensible, a concern that no single metric can operationalize but that every family should be evaluated against. Each of the six family subsections closes with a paragraph explaining its specific connections to the desiderata.

4.1 Group Performance Disparity

This family measures how a scalar performance indicator varies across demographic groups. Let $\mathcal{E}(g)$ denote a task-specific performance measure (e.g., WER, EER, or MOS) for group $g \in \mathcal{G}$. The unified mathematical objective is to quantify the spread of $\{\mathcal{E}(g)\}_{g \in \mathcal{G}}$; different aggregation norms yield different metrics within the same family.

Per-Group Performance Reporting. The most common evaluation practice in the current literature is to report $\mathcal{E}(g)$ for each demographic group without applying an explicit aggregation function. In ASR, per-group WER, Match Error Rate (MER) (Morris et al., 2004), Character Error Rate (CER), and Phoneme

Table 3: Mapping between the fairness desiderata of §2 (Definitions 3–7) and the mathematical families of §4. ● = the family directly operationalizes the definition; ◐ = partial or indirect connection. Definitions 2 and 8 are meta-level constraints discussed in the text.

	Def 3 (DP)	Def 4 (EO)	Def 5 (CF)	Def 6 (Rob.)	Def 7 (Assoc.)
A: Group Perf. Disparity	◐	◐		◐	
B: Statistical Parity	●	●			
C: Calibration & Threshold	◐	●			
D: Representational Disent.	◐		◐		◐
E: Counterfactual Fairness		◐	●		
F: Generative Speech Bias	◐	◐	◐		●

Table 4: Taxonomy of fairness evaluation metrics in speech, organized by mathematical family. Each row groups metrics sharing the same formal objective; the paradigm column indicates the normative affiliation(s) from §3: Robustness (Rob.), Representation (Rep.), or Governance (Gov.).

Math Family	Unified Objective	Representative Metrics	Target Tasks	Paradigm
Group Performance Disparity (§4.1)	Quantify dispersion of $\{\mathcal{E}(g)\}_{g \in \mathcal{G}}$	Δ_{max} , SNSV, Fairness (F), Gini, GARBE, FAAS, Δ_{MOS} ; <i>indiv.</i> : Consistency, L_R	ASR, SV, SER, Diarization, Deepfake, Clinical, TTS	Rob.
Statistical Parity & Cond. Independence (§4.2)	$\hat{Y} \perp G$ (DP) or $\hat{Y} \perp G \mid Y$ (EO)	Δ_{DP} , Δ_{EO} , Equal Opportunity, PPR	SER, Clinical/Health	Rep.
Calibration & Threshold Fairness (§4.3)	Fairness invariance across threshold τ	C_{cal} , ECE, NXE, FaDR, auFaDR-FAR, IR, GARBE	SV, Voice Disorder, Depression/Anxiety	Rob.
Representational Disentanglement (§4.4)	Quantify demographic information in learned representations	MI (DV bound), SpEAT (Cohen’s d), Silhouette/DB/CH indices	SSL Embeddings, SER, SV, Pathological Speech	Rep.
Counterfactual Fairness (§4.5)	$\hat{Y}$ invariance under $do(G)$	ICEO, C3T fairness	ASR, Speech-aware LLMs	Rep.
Generative Speech Bias (§4.6)	Content neutrality w.r.t. demographics	slbs, VoiceBBQ, s_{DIS}/s_{AMB} , $\Delta G/\Delta S$, SAS/BAAS, J/Δ , GUS/SNSR	SLMs, Spoken Dialogue, ST, TTS, Text-to-Audio	Rep.

Error Rate (PER) (Feng et al., 2021) are compared across speaking rates (Mirghafori et al., 2002; Meng et al., 2022b), non-native accents (Veliche & Fung, 2023; Feng et al., 2021; Markl, 2022; Zhang et al., 2023b; 2022d; Maison & Estève, 2023), regional dialects (Tatman, 2017; Veliche & Fung, 2023; Zhang et al., 2022e; Serditova & Tang, 2026), speech impairments (Hu et al., 2024b; Wang et al., 2023b), diverse languages (Emezue et al., 2025; Chauhan & Pundhir, 2026), age (Potamianos & Narayanan, 2003; Feng et al., 2021; Zhang et al., 2024b; Zhao et al., 2023; Pellegrini et al., 2012; Wang et al., 2023b; S & B, 2024; Maison & Estève, 2023), gender (Tatman, 2017; Potamianos & Narayanan, 2003; Sar et al., 2021; Feng et al., 2021; Zhang et al., 2023b; Veliche & Fung, 2023; Garnerin et al., 2019; S & B, 2024; Maison & Estève, 2023), cultural backgrounds (Feng et al., 2021; Zhang et al., 2023b), ethnicity (Koenecke et al., 2020; Veliche & Fung, 2023; Ezema et al., 2025; Si et al., 2021; Lai & Holliday, 2023), disfluency rate (Li & Wu, 2024a), and socioeconomic status (Riviere et al., 2021).

Per-group comparisons extend well beyond ASR. Among speaker-centric tasks, SV reports per-group EER, false match rate (FMR), and false non-match rate (FNMR) (Jin et al., 2022; Fenu & Marras, 2022); speaker diarization reports per-group diarization error rate (DER) (Tevissen et al., 2023); and deepfake detection reports EER and false positive rate (FPR) differentials across gender, age, accent, and language (Singh Yadav et al., 2024b; Staněk et al., 2025; Katamneni et al., 2024). For atypical and clinical speech, per-group reporting covers CER, deletion rate, and diagnostic accuracy gaps (Li & Wu, 2024a; Gulzar et al., 2026; Pahar et al., 2026). The practice even extends to speech synthesis: in dysarthric speech cloning, Parity Difference and Disparate Impact quantify per-severity-level deviations from healthy-speaker baselines (M et al., 2025). Beyond system-side metrics, per-group comparisons extend to the *user-perception* side, uncovering accent-dependent gaps in perceived success and error-recovery cost (Ezema et al., 2025; Michel et al., 2025). Several of the studies above further enhance per-group reporting with significance tests and effect sizes; these statistical validation tools are discussed in § 4.7. While per-group reporting is informative, it leaves the aggregation step implicit; the explicit disparity metrics defined below (Δ_{max} , Δ_{L^2} , Δ_{L^1}) improve cross-study comparability.

Worst-Case Disparity (L^∞). The Maximal Performance Gap (Δ_{max}) formalizes the worst-case pairwise comparison:

$$\Delta_{max} = \max_{g, g' \in \mathcal{G}} |\mathcal{E}(g) - \mathcal{E}(g')|. \quad (14)$$

A large Δ_{max} indicates that at least one pair of groups exhibits a substantial performance gap; the metric therefore prioritizes worst-case equity over average performance. Δ_{max} is widely adopted across ASR fairness evaluations, either in absolute form (Zee et al., 2024; Swain et al., 2026; 2024) or normalized by a reference WER to yield a scale-invariant relative gap suitable for cross-lingual comparison (Attanasio et al., 2024). WER Parity further binarizes this ratio against a tolerance threshold ε for pass/fail auditing (Raes et al., 2026). Alternatively, a *population-referenced* variant measures each subgroup’s deviation from the overall mean, identifying which groups fall above or below the population baseline (Koudounas et al., 2025a). All metrics above index disparity by the speaker’s demographic group; however, the same L^∞ structure can also index the evaluator’s identity. The annotator-gender MOS Gap (Δ_{MOS}), the absolute difference in mean MOS between male and female raters, captures systematic rating bias that propagates into automated MOS prediction models trained on human annotations (Ren et al., 2026).

Variance-Based Disparity (L^2). Variance-based metrics penalize the dispersion of performance across all groups:

$$\Delta_{L^2} = \sqrt{\frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} (\mathcal{E}(g) - \mu_{\mathcal{E}})^2}, \quad (15)$$

where $\mu_{\mathcal{E}}$ is the global mean across all groups. Unlike Δ_{max} , which captures only the worst-case pair, Δ_{L^2} reflects the performance dispersion across the full set of groups. The choice of $\mathcal{E}$ is task-dependent: the Sensitive-to-Sensitive Similarity Variance (SNSV) of Wu et al. (Wu et al., 2026) sets $\mathcal{E} = \text{PRAG}^*@K$ (top- K ranking agreement) to audit spoken dialogue LLMs, while EMO-Debias (Lin et al., 2025c) instantiates as the true positive rate (TPR), FPR, or F1 for SER.

Mean-Deviation Disparity (L^1). The mean-deviation disparity measures the mean absolute deviation (MAD) of group performance:

$$\Delta_{L^1} = \frac{1}{|\mathcal{G}|} \sum_{g \in \mathcal{G}} |\mathcal{E}(g) - \mu_{\mathcal{E}}|. \quad (16)$$

Several domain-specific metrics instantiate Δ_{L^1} with different $\mathcal{E}$: the Fairness score F averages Δ_{L^1} over multiple error rates for biometric verification (Fenu et al., 2020b); FairLENS sets $\mathcal{E} = \text{WER}$ for law-enforcement ASR (Wang et al., 2024b); Overall Bias uses the mean signed WER deviation from a reference group (Patel et al., 2025); and Emo-bias applies Δ_{L^1} to SER accuracy across gender and racial subgroups (Lin et al., 2024d). The Sum of Group Error Differences (SGED) takes a pairwise view, summing $|\mathcal{E}(g) - \mathcal{E}(g')|$ over all group pairs rather than deviations from the mean (Elobaid et al., 2024).

Composite and Welfare-Theoretic Metrics. The ℓ_p metrics above are purely *descriptive*: they quantify how large the disparity is, but not how it should be traded off against overall utility. The metrics below go further by either embedding normative preferences into the aggregation or combining utility and fairness into

a single score. On the normative side, the Gini Coefficient (Gini, 1912) measures inequality via the mean pairwise L^1 distance between group performances, normalized by the overall mean:

$$G = \frac{\sum_{g \in \mathcal{G}} \sum_{g' \in \mathcal{G}} |\mathcal{E}(g) - \mathcal{E}(g')|}{2|\mathcal{G}|^2 \mu_{\mathcal{E}}}. \quad (17)$$

While Δ_{L^1} measures deviations from the mean, G captures the full pairwise structure and is scale-invariant due to the $\mu_{\mathcal{E}}$ denominator. The Isoelastic Social Welfare Function further introduces an inequality-aversion parameter ρ that interpolates between a utilitarian objective and a Rawlsian maximin objective (Triantafyllopoulos & Schuller, 2024). On the composite side, the Gini Aggregation Rate for Biometric Equitability (GARBE), originally developed for face recognition (Howard et al., 2023), aggregates Gini-based inequality across operating points and has since been adapted to SV (Chouchane et al., 2024). For ASR, the Fairness Adjusted ASR Score (FAAS) folds a regression-based fairness penalty into WER to rank systems on leaderboards (Rai et al., 2025). Because these frameworks embed explicit value judgments, researchers should make their parameter choices transparent when reporting results.

Individual-Level Analogues. All preceding metrics index disparity by demographic group $g \in \mathcal{G}$; individual fairness instead requires that *similar samples receive similar predictions*, regardless of group membership (Dwork et al., 2012). The same ℓ_p norm structure extends naturally to this setting. The Consistency Score (Zemel et al., 2013) measures the average L^1 deviation between each sample’s prediction and the mean prediction of its k -nearest neighbors in embedding space, mirroring Δ_{L^1} at the sample level:

$$\text{Consistency} = 1 - \frac{1}{n} \sum_{i=1}^n \left| \hat{y}_i - \frac{1}{k} \sum_{j \in \text{kNN}(i)} \hat{y}_j \right|, \quad (18)$$

where the ideal value is 1; this metric is used in SER (Chien & Lee, 2023; 2024). The Lipschitz constant $L_R = \sup_{x \neq x'} |f(x) - f(x')| / d(x, x')$ captures worst-case prediction sensitivity, analogous to Δ_{max} but normalized by input distance; a lower L_R indicates smoother decisions between acoustically similar utterances (Chien & Lee, 2026). Empirical evidence in SER shows a trade-off: optimizing group-level parity can degrade individual-level consistency, particularly when attribute distributions between groups are distant (Chien & Lee, 2024; 2026).

Connection to Fairness Desiderata. This family quantifies the *spread* of $\mathcal{E}(g)$ across groups rather than testing a distributional independence condition, so the connections to Def. 3 and 4 are both partial. Def. 4 (EO, $\bullet$): because $\mathcal{E}(g)$ is usually computed against the ground truth (e.g., WER, EER), a nonzero gap often signals an EO violation when $P(Y | G)$ is similar across groups, but a gap can also arise from differing base rates under perfect EO. Def. 3 (DP, $\bullet$): when $\mathcal{E}$ is an unconditional statistic such as the overall acceptance rate the test is closer to a DP check, though most instantiations in the literature use ground-truth-referenced error metrics, limiting this reading to a subset of use cases. Def. 6 (Robustness, $\bullet$): the metrics of this family capture performance gaps under naturalistic variability, while Def. 6’s stress-test criterion is operationalized by robustness benchmarks that compare error-rate degradation across demographic groups under injected noise, channel, and codec distortions (Rajan et al., 2022; Shah et al., 2025; Baali et al., 2025).

Limitations. All metrics in this family treat demographic identity as an extraneous source of variance to minimize, without questioning whether the performance measure $\mathcal{E}$ itself is culturally appropriate or whether the group taxonomy $\mathcal{G}$ captures the relevant social categories. A narrow inter-group gap does not imply fairness if absolute performance remains poor for all groups, or if the evaluation metric fails to reflect the construct it purports to measure (see Def. 8).

4.2 Statistical Parity and Conditional Independence

Moving beyond the scalar performance comparisons of § 4.1, this family tests whether the model’s *prediction distribution* $P(\hat{Y} | G)$ is independent of the sensitive attribute G . The gap metrics there reduce each group to one aggregate score, fusing per-class behavior with the group’s base rates $P(Y | G)$. DP instead requires $\hat{Y} \perp G$ while ignoring Y ; EO requires the same within each true class. Aggregate gaps and these independence tests therefore need not agree. Per-class disparities may cancel out in the aggregate, and even under perfect EO, differing base rates alone can produce a nonzero gap.

Core Metrics: Δ_{DP} and Δ_{EO} . For a classification task with predicted output $\hat{Y} \in \mathcal{Y}$ and sensitive attribute G , the per-class gap between two groups is:

$$d_{DP}(y, g, g') = |P(\hat{Y}=y \mid G=g) - P(\hat{Y}=y \mid G=g')|. \quad (19)$$

Demographic Parity (DP) requires this gap to vanish for all classes and group pairs. When base rates naturally vary across groups (e.g., in SER, where emotional expressiveness differs by culture), enforcing DP can penalize legitimate variance (Hardt et al., 2016; Dwork et al., 2012). Equalized Odds (EO) addresses this by conditioning on the ground-truth label Y :

$$d_{EO}(y, g, g') = |P(\hat{Y}=y \mid G=g, Y=y) - P(\hat{Y}=y \mid G=g', Y=y)|, \quad (20)$$

requiring equal accuracy across groups *within* each true class, thereby preserving natural variance while prohibiting demographic shortcuts (Hardt et al., 2016; Lin et al., 2025b). To obtain a scalar summary Δ_{DP} (or Δ_{EO}), these per-class gaps must be aggregated over $\mathcal{Y}$ and $\mathcal{G}$. Common choices include worst-case ($\max_y \max_{g, g'}$), macro-averaging over classes ($\frac{1}{|\mathcal{Y}|} \sum_y \max_{g, g'}$), and restricting to the positive class only (Equal Opportunity, a special case of EO) (Lin et al., 2025c; Haghighi et al., 2026). The choice of aggregation affects sensitivity: worst-case highlights the single most unfair (class, group) pair, while macro-averaging reflects disparity across the full label space.

Domain Applications. In SER, macro-averaged Δ_{EO} and Δ_{DP} jointly quantify the trade-off between debiasing and accuracy (Lin et al., 2025c); cross-corpus evaluations use Δ_{DP} and Δ_{EO} to test whether gender fairness transfers across datasets (Upadhyay et al., 2025); and dual-constraint frameworks study how Δ_{EO} trades off against individual-level consistency (Chien & Lee, 2026; 2024). In clinical settings, the predictive positive rate (PPR) and Statistical Parity deconstruct biases across intersecting demographic factors in health predictions (Yang et al., 2024), while Equal Opportunity and Equalized Odds are applied to multi-class cognitive impairment detection across gender, age, education, and language subgroups (Haghighi et al., 2026).

Connection to Fairness Desiderata. This family directly operationalizes both Def. 3 (Demographic Parity, ●) and Def. 4 (Equalized Odds, ●). Δ_{DP} tests $\hat{Y} \perp G$, which is exactly the constraint stated in Def. 3; Δ_{EO} tests $\hat{Y} \perp G \mid Y$, matching Def. 4. The two metrics correspond one-to-one with the mathematical statements in Def. 3 and 4, making this family the most direct bridge between the formal desiderata and computable evaluation.

Limitations. Both Δ_{DP} and Δ_{EO} are defined over discrete predicted classes; extending them to continuous outputs such as MOS or confidence scores requires discretization, and empirical evidence shows that fairness conclusions can reverse when the operating point shifts (Hutiri et al., 2024b). Moreover, Δ_{EO} conditions on the ground-truth label Y . Any bias present in the annotations propagates directly into the fairness assessment. In SER, where labels reflect annotator subjectivity and speaker-rater demographic interactions (Chien et al., 2024), apparent compliance with EO may mask rather than resolve unfairness (see § 5.1.3).

4.3 Calibration and Threshold Fairness

§ 4.1 and § 4.2 evaluate disparity at a fixed operating point. This family treats the decision threshold τ as a *free variable*, asking whether a fairness conclusion reached at one operating point still holds at another. Two complementary perspectives arise: *score calibration*, which tests whether output scores carry the same probabilistic meaning across groups regardless of τ , and *threshold sweep*, which measures fairness as a function of τ across the full operating range.

Score Calibration. The cost of log-likelihood ratios (C_{llr}) decomposes into a discrimination term and a calibration term C_{cal} , where C_{cal} isolates how well output scores are calibrated independently of the system’s ability to separate target from non-target trials (Brümmer & du Preez, 2006). In SV, C_{cal} degrades dramatically for underrepresented accents even when discrimination performance remains stable (Estevez & Ferrer, 2023). In voice disorder detection, Normalized Cross-Entropy (NXE) reveals similar calibration gaps across gender and age subgroups, which group-dependent Platt scaling substantially reduces (Estevez et al., 2025). In speech-based depression and anxiety prediction, Expected Calibration Error (ECE) likewise

varies across demographic groups (Norbury et al., 2026). Across all three tasks, the pattern is consistent: discrimination and calibration are separable, and a system can rank speakers correctly yet produce scores whose confidence is systematically misleading for certain populations.

Threshold Sweep. The Fairness Discrepancy Rate $\text{FaDR}(\tau) = 1 - [\alpha \cdot \Delta_{\text{FMR}}(\tau) + (1 - \alpha) \cdot \Delta_{\text{FNMR}}(\tau)]$ (de Freitas Pereira & Marcel, 2022) combines the largest inter-group FMR gap (Δ_{FMR}) and FNMR gap (Δ_{FNMR}) at each threshold, with $\text{FaDR} = 1$ indicating perfect fairness. Integrating FaDR over the FMR axis yields a threshold-free summary, auFaDR-FAR (Peri et al., 2022). FaDR , IR (Grother, 2022), and GARBE have been compared head-to-head for SV (Chouchane et al., 2024), using the functional fairness measure criteria of Howard et al. (Howard et al., 2023) as the yardstick. These criteria ask whether the contributions of the two error rates remain intuitive, whether the score stays within defined bounds, and whether it remains computable when a subgroup achieves zero errors. Only GARBE satisfies all three.

Connection to Fairness Desiderata. Def. 4 (EO, $\bullet$): FaDR , auFaDR-FAR , and GARBE compare group-level FMR and FNMR at each threshold τ , directly testing $P(\hat{Z}=1 \mid Z=z, G=g) = P(\hat{Z}=1 \mid Z=z, G=g')$ across the full operating range. The score-calibration metrics (C_{cal} , ECE , NXE) complement this by testing whether the score carries the same probabilistic meaning across groups ($P(Y=1 \mid S=s, G=g) = P(Y=1 \mid S=s)$), which supports but does not formally entail EO. Def. 3 (DP, $\bullet$): threshold-sweep analysis may incidentally reveal that $P(\hat{Z}=1 \mid G=g)$ diverges across groups at certain operating points, but the family’s primary focus is conditional fairness.

To date, these metrics have been developed almost exclusively for SV and voice-disorder detection; extending them to other speech tasks remains open.

4.4 Representational Disentanglement

The preceding families evaluate bias at the output level. This family moves upstream, quantifying how much demographic information leaks into learned speech representations before any decision is made.

The central metric is Mutual Information ($MI(Z; G)$) between the embedding Z and the sensitive attribute G : a high value indicates that the representation encodes demographic identity, enabling downstream classifiers to exploit group membership even when the primary task is unrelated to demographics. Because exact MI is intractable for high-dimensional speech vectors, practical estimation relies on variational lower bounds such as the Donsker–Varadhan (DV) bound, which trains a neural discriminator to tighten a tractable approximation. In the speech domain, WavShape applies this strategy to self-supervised learning (SSL) embeddings (Baser et al., 2025).

Beyond MI, several complementary evaluation strategies have emerged. The Speech Embedding Association Test (SpEAT) adapts the Implicit Association Test from psychology to speech embeddings (Slaughter et al., 2023). It quantifies, via Cohen’s d , the differential association between two groups of speaker stimuli (e.g., male vs. female voices) and two sets of evaluative attributes (e.g., pleasant vs. unpleasant); applied to major self-supervised architectures, SpEAT reveals that these models amplify social biases across gender, age, and nationality relative to classic acoustic features (Lin et al., 2024b). From a geometric perspective, clustering quality indices such as the Silhouette score (Rousseeuw, 1987), Davies–Bouldin index (Davies & Bouldin, 1979), and Calinski–Harabasz index (Caliński & Harabasz, 1974) can quantify how tightly embeddings group by demographic category rather than by task-relevant class. In pathological speech detection, these indices reveal the degree to which accent variation confounds clinical markers across multilingual populations (Kashyap & Pathirana, 2026). On the statistical side, mixed-effects regression on ℓ_2 -normalized embedding distances separates the variance attributable to group membership from that attributable to channel or phonetic content in speaker verification (Jahan et al., 2023).

Connection to Fairness Desiderata. Definition 3 (DP, $\bullet$): if $MI(Z; G) = 0$, then $Z \perp G$, and any downstream prediction $\hat{Y} = h(Z)$ must also satisfy $\hat{Y} \perp G$ by the data-processing inequality. Therefore, disentanglement provides an upstream sufficient condition for DP, though the connection is approximate in practice because the DV bound yields only a lower bound on MI. Def. 5 (Counterfactual Fairness, $\bullet$): SpEAT asks whether group membership produces asymmetric associations in embedding space, which relates to Def. 5’s output invariance requirement, but the connection is indirect because SpEAT quantifies association

asymmetry rather than invariance under an explicit transformation $T_{g \rightarrow g'}$. Def. 7 (Output Association Equality, $\bullet$): when applied to generative speech models, SpEAT’s association measurement overlaps with Def. 7’s $s(\bar{Y})$, but operates in embedding space rather than on the final output.

Limitation. These metrics share two important limitations. First, the DV bound provides only a lower bound on MI, so a value near zero does not guarantee that the representation is free of demographic information; the true leakage may be higher. Second, low demographic information in the representation does not ensure downstream fairness: a large-scale evaluation of SER shows that upstream representational bias and downstream performance disparity can diverge (Lin et al., 2024d).

4.5 Counterfactual Fairness

The preceding families evaluate fairness at the group level; this family adopts a causal perspective: changing a speaker’s demographic attribute while holding linguistic content constant should not alter the model’s output (Kusner et al., 2017). In speech, operationalizing this principle requires a perturbation mechanism that modifies perceived demographics without changing lexical content.

For ASR, Sari et al. formalize the Individualized Counterfactual Equal Odds (ICEO) criterion, which requires that outcome probabilities remain invariant when a speaker’s demographic attributes are counterfactually altered; in practice, compliance is evaluated by measuring CER variation across individuals on counterfactually augmented speech (Sar et al., 2021). For speech-aware LLMs, the Cross-modal Capabilities Conservation Test (C3T) benchmark defines a counterfactual fairness metric as $\text{fairness} = \# \text{fair_tasks} / \# \text{tasks}$, where a task is fair if it yields the same answer for every speaker regardless of demographics; conditional variants further decompose this score by specific attributes such as age, gender, and dialect (Kubis et al., 2025).

Connection to Fairness Desiderata. This family directly operationalizes Definition 5 (Counterfactual / Invariance Fairness, $\bullet$). Both ICEO and C3T’s fairness metric formalize the requirement that $M_\theta(X)$ remains invariant when the speaker’s demographic attributes are counterfactually altered via $T_{g \rightarrow g'}$, which is the defining equation of Def 5. Definition 4 (Equalized Odds, $\bullet$) is partially reflected. ICEO is explicitly named “Individualized Counterfactual *Equal Odds*,” embedding the EO constraint within a causal framework: it requires not only that error rates be equal across groups, but that this equality hold at the individual level under counterfactual demographic change.

Limitation. Counterfactual fairness evaluation in speech remains an open direction: the perturbation mechanisms available today, including adversarial auto-encoders and voice conversion, inevitably alter acoustic properties beyond the target demographic attribute, making it difficult to isolate causal effects cleanly.

4.6 Bias in Generative Speech Systems

Mathematically, most metrics in this section build on the parity and proportion comparisons introduced in § 4.2, though several adopt alternative formulations such as set overlap, log-odds interaction, or composite scoring. They are grouped separately because the evaluation target is fundamentally different: whereas § 4.1–4.2 compare system *performance* across demographic groups of input speakers, the metrics here evaluate whether the *content* produced by generative models, including Speech Language Models (SLMs), speech translation (ST), TTS, and text-to-audio synthesis, encodes social stereotypes or demographic preferences. Recent holistic benchmarks such as AudioTrust (Li et al., 2026) and AHELM (Lee et al., 2025a) treat fairness as one of several trustworthiness dimensions for audio language models, each introducing its own composite scoring. In parallel, a structural taxonomy of audio LLMs highlights the need for architecture-aware evaluation across cascaded, multimodal, and audio-native designs (Aloufi et al., 2026).

SLMs and Spoken Dialogue. At the language-modeling level, the Speech Language Bias Score (slbs) measures the percentage of instances in which a model prefers a stereotypical over an anti-stereotypical continuation (Lin et al., 2024a). In question answering, a key design axis is whether the context uniquely determines the correct answer (disambiguated) or leaves it open (ambiguous). VoiceBBQ reports accuracy on disambiguated items and a bias score on ambiguous items, additionally separating content bias from acoustic bias (Choi et al., 2025a). Listen and Speak Fairly independently adopts the same split with

$s_{\text{DIS}} = 2(n_{\text{biased}}/n_{\text{not_unknown}}) - 1$ and $s_{\text{AMB}} = (1 - \text{accuracy}) \times s_{\text{DIS}}$ (Lin et al., 2024c). Orthogonal to textual content, Gender Response Overlap (J) measures the distributional overlap between responses to male and female voices, while Gender Preference (Δ) captures how frequently a model’s response aligns with a specific gender (Choi et al., 2025b). At the downstream decision level, the Group Unfairness Score (GUS) and Similarity-based Normalized Statistics Rate (SNSR) measure decision-level and recommendation-level inequity across demographic groups in spoken dialogue systems (Wu et al., 2026). To avoid the expressiveness constraints of multiple-choice question (MCQ) formats, which restrict stereotypical associations to predefined options, VIBE (Lin et al., 2026b) evaluates 11 SLMs through open-ended generation tasks with real-world (not synthetic) speech, measuring distributional shifts via normalized Total Variation Distance (nTVD); it reveals systematic biases undetected by MCQ benchmarks. Biased conclusions can also be sensitive to the evaluation format itself. The Average Pairwise Entropy Shift (APES) quantifies this sensitivity by averaging pairwise entropy differences across levels of a controlled variable. Paired with Fleiss’ κ , it shows that multilingual MLLMs are relatively robust to gender and accent but highly sensitive to language and option order, with speech systematically amplifying text-modality biases (Wei et al., 2026). A caveat is that bias scores obtained from multiple-choice benchmarks do not reliably transfer to open-ended generation, so conclusions drawn from a single evaluation format should be interpreted with caution (Satish et al., 2026a).

Speech Translation. Translating from a language without grammatical gender, such as English, into gender-marked languages such as Spanish or Italian requires the system to infer the gender of each entity. Systems that default to masculine forms or rely on stereotypical gender-occupation associations exhibit lower accuracy on female and anti-stereotypical instances. The WinoST challenge set quantifies this effect through gender accuracy, the proportion of correctly gendered entities, alongside ΔG , the gap between male and female F1, and ΔS , the gap between pro-stereotypical and anti-stereotypical F1 (Costa-jussà et al., 2022; Lin et al., 2024c). MuST-SHE complements this evaluation by computing accuracy on both correct and gender-swapped references; the difference between the two scores reveals the direction of bias, exposing whether the system systematically favors masculine over feminine forms (Bentivogli et al., 2020).

Speech and Audio Synthesis. A common evaluation strategy for instruction-guided TTS estimates the gender probability of synthesized speech via a pretrained speaker gender recognition classifier and compares it against expected distributions (Kuan & Lee, 2025). When instructions combine multiple social dimensions, the interaction term $\mathcal{I}$, defined as the log-odds deviation from an additive baseline, isolates how status, career, and persona cues jointly modulate gender probability beyond their individual effects (Chen et al., 2026a). In unprompted speech generation, the female utterance ratio across repeated syntheses of the same text quantifies systematic gender preferences in automatic speaker assignment (Puhach et al., 2025). For text-to-audio generation, the Stereotype Audio Score (SAS), defined as $\max(\%N_{\text{male}}, \%N_{\text{female}})$, captures the gender dominance in a pool of generated samples, and the Bias-Aware Audio Score (BAAS) combines SAS with an audio quality term to penalize high-fidelity generation that lacks gender balance (Mohsin et al., 2025).

Connection to Fairness Desiderata. Def. 3 and 4 (DP and EO, $\bullet$): GUS and SNSR test whether decision outcomes (recommendations, system actions) differ across groups, mapping to DP/EO at the decision level rather than at the content level. Def. 5 (Counterfactual Fairness, $\bullet$): VoiceBBQ’s content-vs-acoustic decomposition asks whether changing the speaker’s demographic attributes alters the model’s response, mirroring the counterfactual invariance of Def. 5. Def. 7 (Output Association Equality, $\bullet$): slbs, VoiceBBQ, $s_{\text{DIS}}/s_{\text{AMB}}$, $\Delta G/\Delta S$, and SAS/BAAS each test whether a derived output attribute (stereotype score or gender probability) is independent of G , instantiating a specific choice of s in Def 7’s $\mathbb{E}[s(\hat{Y}) \mid G=g] = \mathbb{E}[s(\hat{Y}) \mid G=g']$.

4.7 Statistical Validation

The preceding families quantify the magnitude of demographic disparities. This subsection collects the statistical tools needed to determine whether those disparities are genuine rather than artifacts of sampling variability.

Any observed disparity may arise from sampling noise rather than systematic bias, particularly for small or intersectional subgroups. Significance tests such as ANOVA establish whether group differences exceed chance expectations, while effect-size measures such as Cohen’s d (Cohen, 2013) quantify their practical magnitude.

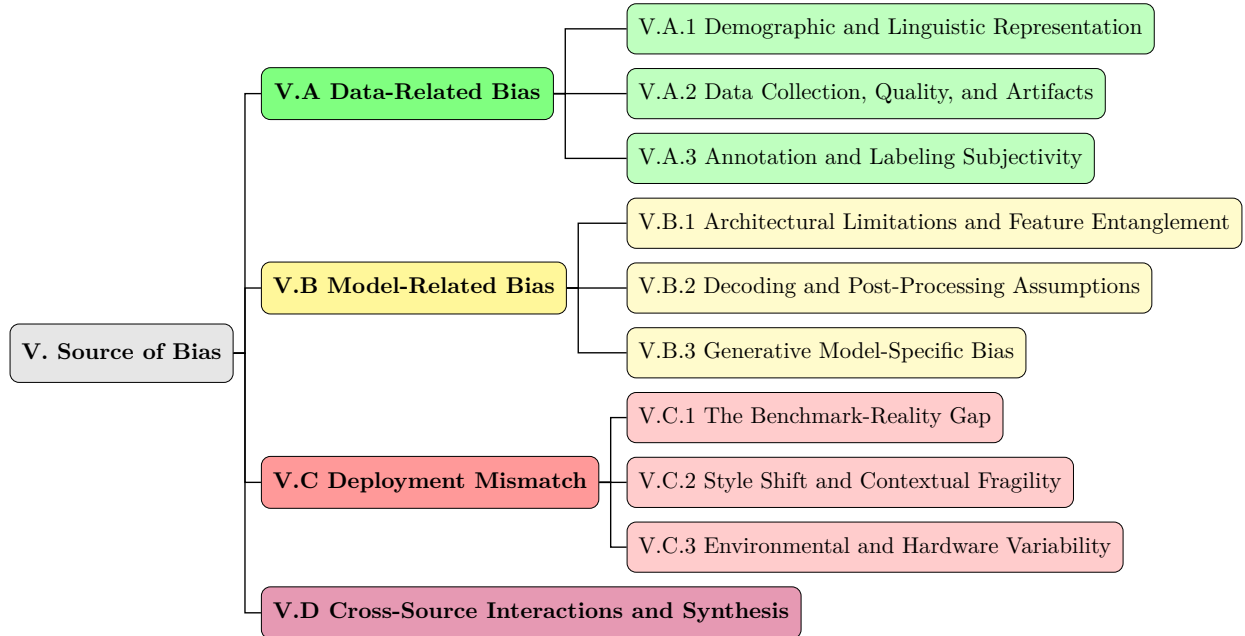


Figure 2: Taxonomy of bias sources in speech technology.

When many subgroups are compared simultaneously, the Benjamini-Hochberg procedure controls the false discovery rate (Benjamini & Hochberg, 1995; Altwlkany et al., 2025). For the skewed error-rate distributions common in speech, non-parametric alternatives such as the Mann-Whitney U (Mann & Whitney, 1947) and Wilcoxon signed-rank tests (Wilcoxon, 1945) offer robust unpaired and paired comparisons (Puhach et al., 2025; Székely et al., 2023). When confounding variables such as recording conditions or annotator subjectivity are present, mixed-effects regression isolates the true demographic contribution by treating group attributes as fixed effects and nuisance factors as random effects, with Poisson or logistic link functions for count-based and binary outcomes (Jimenez & Kern, 2026; Rai et al., 2025; He et al., 2026). Finally, utterances from the same speaker are statistically dependent, violating the independence assumption of the above tests; blockwise bootstrap resampling at the speaker level preserves this correlation structure and yields more reliable confidence intervals (Liu & Peng, 2020).

Beyond the quantifiable disparities addressed throughout this section, the Governance paradigm encompasses non-quantifiable concerns such as data sovereignty and community agency (§ 3.3), which require institutional rather than algorithmic interventions.

5 Source of Bias

As illustrated in Figure 2, bias in speech processing systems stems from three primary sources: *data-related issues*, *model-related factors*, and *deployment mismatch*. These three sources are complementary to, but not identical to, the conceptual paradigms introduced in § 3; they are organized by point of origin in the pipeline rather than by normative concern, and the two taxonomies intersect freely. We detail each source below and discuss their interactions in § 5.4.

5.1 Data-Related Bias

Data-related biases are foundational, arising either from skewed data distributions or inherent differences in speech characteristics that are inadequately represented.

5.1.1 Demographic and Linguistic Representation

Biases in training data generally stem from two primary deficits: imbalances in the *demographic representation* of speaker groups and insufficient coverage of *diverse linguistic content*.

Demographic Underrepresentation. Standard datasets exhibit imbalances that reflect societal hierarchies rather than population distributions. Along the *gender* axis, broadcast news and web corpora favor male speakers, yielding systems less robust to female acoustic patterns (Garnerin et al., 2019; Maison & Estève, 2023; Lin et al., 2025b). Along the *age* axis, adult-centered distributions leave child speech (whose shorter vocal tracts produce distinct pitch and formant patterns) and elderly speech severely underrepresented (Potamianos & Narayanan, 2003; Pellegrini et al., 2012; Borre et al., 2025). *Racial and ethnic* imbalances are mediated by the marginalization of non-standard ethnolects: AAE speakers suffer systematic transcription errors when evaluated against corpora dominated by Standard American English (Koencke et al., 2020; Martin & Wright, 2023; Brewer et al., 2023; Wenzel et al., 2023). Beyond demographic axes, *atypical speech* is systematically excluded, as standard datasets prioritize fluent speech and leave conditions such as stuttering, dysarthria, and Parkinson’s-affected speech scarcely represented (Hu et al., 2024b; Li & Wu, 2024a; Singh Yadav et al., 2024a; Wang et al., 2023c; Moro-Velazquez et al., 2019). Finally, *metadata gaps* constitute a distinct representational harm: the lack of granular labels for non-binary gender identities and sociophonetic markers prevents fairness evaluation for queer speakers (Sigurgeirsson & Ungless, 2024).

Linguistic and Content Disparities. Beyond demographic labels, most training corpora implicitly privilege standard language varieties, reflecting the standard language ideology that elevates prestige dialects as the normative baseline while treating other variations as deviations (Humphries et al., 1996; Markl, 2022; Chen et al., 2025b). Unlike acoustic mismatches, this bias manifests in the language modeling layer, where valid dialectal syntax is misinterpreted as errors because training data enforces standard grammatical rules. Corpora also lack diversity in cultural and idiomatic content, producing systematically higher error rates for subgroups whose linguistic habits diverge from the training norm (Koudounas et al., 2025a).

5.1.2 Data Collection, Quality, and Artifacts

Bias arises from systematic inequalities in both the *physical* and *procedural* aspects of data creation, alongside the hidden artifacts introduced during the collection pipeline.

Socioeconomic and Equipment Disparities. Socioeconomic disparities manifest as *channel bias*, a form of *Technical Bias* (§ 2.1): data from marginalized populations is disproportionately captured in high-reverberation environments or on low-fidelity hardware, so models may use channel characteristics as a demographic proxy instead of learning robust linguistic features (Papakyriakopoulos & Xiang, 2023). This confound can arise even within a single corpus, where recording practices co-vary with speaker subgroup (Li et al., 2024). Crowdsourced datasets exacerbate the issue, as minority-group acoustic conditions diverge from studio-quality training data; speech enhancement preprocessing can further introduce artifacts that disproportionately degrade lower-resource recordings (Giraldo et al., 2025).

Acoustic and Signal Variability. Beyond equipment, bias resides in latent acoustic features that demographic counts alone cannot capture. Spectral analysis and learned representations are often optimized for the lower fundamental frequencies typical of adult males, leaving systems ill-suited to the pitch and spectral patterns of other speakers (Tevissen et al., 2023; Koencke et al., 2020; Gorrostieta et al., 2019). More subtly, imbalances in speech rate (Meng et al., 2022b), prosody, and rhythm (Baser et al., 2025; Lai & Holliday, 2023) can drive stronger bias than demographic labels themselves (Gu et al., 2023), acting as hidden confounders that persist undetected even in demographically balanced datasets because they are rarely logged in metadata.

Collection Protocols and Transcription. Recording and labeling protocols systematically place different groups in unequal conditions. In broadcast corpora, women are more often recorded in formal scripted settings while men appear in spontaneous conversational contexts, producing spurious correlations between gender and signal quality (Markl & McNulty, 2022); more broadly, datasets favoring read speech yield optimistic performance estimates that collapse under real-world conditions (Patel et al., 2025; Raes et al., 2026). Transcription guidelines are similarly non-neutral: choices regarding disfluencies, non-standard grammar,

and hesitation markers disproportionately inflate error rates for speakers of non-standard varieties (Feng et al., 2021; Reitmaier et al., 2022), while protocols demanding sustained, fluent articulation exclude medical subgroups whose speech is affected by conditions such as medication-induced articulation changes (Teodorescu et al., 2023).

Metadata Gaps and Data Artifacts. Bias is also embedded in the infrastructure of datasets. Audits of open-source corpora reveal that essential metadata, such as gender demographics and speech duration, is frequently missing or inconsistent, leaving fairness claims about these datasets untestable (Garnerin et al., 2019). Moreover, dataset construction functions as a mechanism for setting norms: the varieties included and documented determine whose speech is treated as *normal* by downstream systems, while undocumented varieties are implicitly categorized as deviations (Markl & McNulty, 2022). These gaps enable shortcut learning, where systems rely on dataset-specific artifacts rather than genuine linguistic features, so that strong benchmark performance may not reflect true robustness (Palmer et al., 2021; Papakyriakopoulos & Xiang, 2023).

5.1.3 Annotation and Labeling Subjectivity

Bias in annotation arises because speech labels are judgments filtered through human labelers’ subjective expectations, shaped by both *cognitive priors* in individual perception and *cultural distance* between rater and speaker.

Perceptual and Cognitive Bias. Human perception introduces non-random label noise shaped by cognitive priors independent of the signal. Raters exhibit systematic valence biases that vary with age and modality (Kauschke et al., 2019), and stronger implicit bias correlates with amplified neural responses to foreign-accented speech, shifting evaluations of clarity, competence, or affect regardless of content (Yi et al., 2014). This produces an annotator-speaker mismatch in which subgroup differences in labels emerge not from the signal but from how raters interpret identical cues across demographics (Chien et al., 2024; Chien & Lee, 2023). Such systematic label errors are fossilized as ground truth, so that downstream models learn and scale annotators’ implicit prejudices.

In-Group Advantage and Cultural Distance. Labeling reliability degrades as cultural distance between rater and speaker increases: annotators show higher accuracy and agreement for speakers from their own demographic background and lower accuracy for accents they do not share (Papakyriakopoulos & Xiang, 2023). This distance leads raters to misclassify valid dialectal variations as ambiguity or errors, injecting systematic noise correlated with speaker identity (Feng et al., 2021). Bias also enters the semantic content of labels: gender stereotypes, such as rating female speech as more emotional, become encoded in training labels and amplified by deployed systems (He et al., 2026; Rizhinashvili et al., 2022).

5.2 Model-Related Bias

Even with balanced training data, bias can arise from model-intrinsic design choices, manifesting as *architectural limitations* in learned representations, *deterministic assumptions* embedded in decoding logic, and mechanisms specific to *generative speech models*.

5.2.1 Architectural Limitations and Feature Entanglement

Representation Entanglement. Standard optimization objectives in tasks such as speaker verification often fail to disentangle speaker identity from linguistic content and environmental factors, causing speaker embeddings to encode language-specific phonological information and misinterpret language shifts as identity shifts (Sharma, 2024). This entanglement has demographic consequences: gender-related performance gaps persist across standard speaker verification architectures (Hutiri & Ding, 2022).

Tokenization and Vocabulary Bias. Tokenization is an often-overlooked bias source in end-to-end speech models that jointly learn acoustic and linguistic representations. Byte-pair encoding tokenizers in models such as Whisper (Radford et al., 2023) are trained predominantly on high-resource languages, yielding higher token likelihoods for those languages and worse performance on low-resource ones (Liang et al., 2025). The resulting over-segmentation forces decoding over longer, less informative token sequences for underrepresented

languages (Tripathi et al., 2025), and compounds in code-switching, where models degrade at language-switch boundaries despite strong per-language performance (Liu et al., 2026). Because the bias lies in the learned vocabulary rather than the training audio, more speech data cannot remedy it: sub-token utilization is shaped more by typological and orthographic structure than by training scale (Liang et al., 2026), so interventions must act at the tokenization layer.

Model Compression and Capacity Bias. Compression techniques such as pruning, quantization, and knowledge distillation can introduce or amplify demographic bias even when applied to a fair base model. Pruning disproportionately degrades performance on underrepresented, atypical long-tail examples while leaving aggregate accuracy largely unaffected (Hooker et al., 2019). In speech, this effect is further modulated by architectural shape: the trade-off between model width and depth influences bias more than raw parameter count (Lin et al., 2024b). Fairness audits must therefore extend beyond pre-deployment evaluation of the full model to include post-compression testing, since compression itself is a bias-introducing design choice.

5.2.2 Decoding and Post-Processing Assumptions

Hard-coded decoding heuristics apply uniformly across deployment contexts, unlike the context-dependent failures of § 5.3. Voice activity detection (VAD) thresholds misclassify dysfluent pauses as silence, and language models trained on standard fluency probabilities over-correct repetitions; both disadvantage speakers who stutter (Lea et al., 2023). The same assumptions hurt users with cognitive conditions: individuals with dementia need mid-utterance pauses for cognitive planning, but turn-taking algorithms misread these as end-of-turn signals and cut off the user (Addlesee & Eshghi, 2024). The burden falls on atypical speakers, compounding the data underrepresentation identified in § 5.1.1 (Li & Wu, 2024a).

5.2.3 Generative Model-Specific Bias

Unlike discriminative models, generative systems such as TTS, Voice Conversion (VC), spoken dialogue, and SLLMs actively *construct* speaker identities, creating new bias categories. When no speaker is specified, TTS and spoken dialogue systems default to voices aligned with dominant social groups, encoding occupational gender stereotypes, cultural appropriation, and identity erasure into the output (He et al., 2026). Autoregressive architectures may further amplify such biases by propagating errors from early decoding steps, though the differential demographic impact remains an open research question.

5.3 Deployment Mismatch

Deployment mismatch, corresponding to *Emergent Bias* in § 2.1, arises when systems validated on idealized benchmarks fail in contexts that differ significantly from their design environment.

5.3.1 The Benchmark-Reality Gap

Sterile benchmarks mask shortcut learning: developers optimize for metrics decoupled from operational reality, allowing models to exploit dataset artifacts such as recording devices or background noise as spurious proxies for prediction targets. In clinical speech assessment, models may associate microphone frequency responses with cognitive decline scores and collapse when deployed with standardized hardware (Low et al., 2024). At production scale, audits of ASR systems reveal cohort-level performance disparities invisible in benchmark evaluation (DHERAM et al., 2022).

5.3.2 Style Shift and Contextual Fragility

Rigid domain assumptions create barriers when systems trained on scripted speech face spontaneous settings. Elderly users introduce physiological artifacts (tremors, breathiness, cognitive planning pauses) largely absent from standard corpora (Vipperla et al., 2010). For dialect speakers, dynamic style shifting between prestige varieties and regional vernaculars introduces phonetic and prosodic variations that produce systematic recognition failures when models are trained only on careful, isolated speech (Markl, 2022).

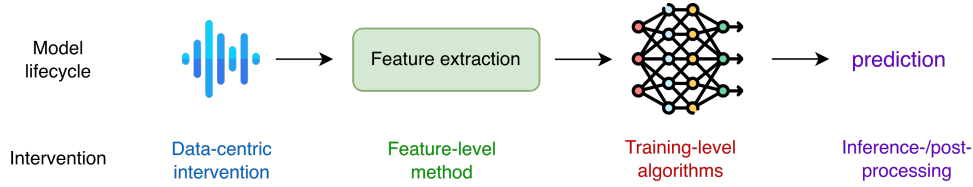


Figure 3: Intervention methods across the speech model production pipeline.

5.3.3 Environmental and Hardware Variability

Beyond channel bias in training data, deployment environments introduce variability through real-world noise and heterogeneous hardware that interact with speaker demographics.

Noise and Demographic Interaction. Background noise degrades ASR performance unevenly across speaker cohorts: the acoustic distance between a speaker’s pronunciation and the model’s learned prototypes widens under noise, and this widening is systematically greater for speakers already at the margins of the training distribution (DHERAM et al., 2022). Common preprocessing can exacerbate this; speech enhancement frontends may introduce artifacts that disproportionately affect underrepresented speech patterns, producing a counterproductive cycle where denoising hurts the populations it should help (Giraldo et al., 2025).

Device Heterogeneity. Users interact with speech systems through diverse hardware (studio microphones, far-field smart speakers, budget smartphones) that impose distinct acoustic transfer functions. Emerging evidence indicates hardware can interact with demographics (e.g., contact microphone placement produces gender-dependent accuracy gaps (Konuma & Asakura, 2023)), but systematic audits across device types remain largely absent.

5.4 Cross-Source Interactions and Synthesis

While the preceding subsections present data, model, and deployment biases as analytically distinct categories, real-world fairness failures typically involve their interaction, producing compounding effects that exceed the sum of individual contributions. We identify three such interaction patterns.

Cascading Pipeline Effects. In speech systems, bias introduced at an upstream stage propagates to and is amplified by downstream components in any multi-stage pipeline, turning acoustic-level failures into semantic-level ones (Ezema et al., 2025).

Intersectional Compounding. Bias that appears moderate along a single demographic axis can compound severely at demographic intersections (Harris et al., 2024). Single-attribute fairness audits therefore systematically underestimate the harm borne by those at multiple marginalized intersections.

Architecture and Data Amplification. Model design choices can selectively amplify latent data biases rather than merely reflect them (Hutiri & Ding, 2022; Hooker et al., 2019). The final fairness profile of a deployed system is thus not reducible to its training data alone, but is jointly determined by the interaction between data composition and architectural decisions.

6 Mitigation Strategy

Research on fairness in speech technologies encompasses a broad spectrum of interventions. We organize them by their point of action in the model production pipeline (Figure 3), yielding four stages. **Data-centric interventions** act before any model is trained, reshaping the input distribution to address historical underrepresentation. **Feature-level methods** act on representations that have already been learned, transforming embeddings with the encoder weights held fixed. **Training-level algorithms** act during optimization, modifying loss functions, gradient signals, or architectural pathways so that fairness is enforced as representations are formed. **Inference- and post-processing strategies** act after deployment, adjusting decisions without altering either the data or the model. We draw the boundary between the feature level

and the training level by whether the encoder is modified during debiasing: feature-level methods transform representations extracted by a fixed encoder, whereas training-level methods reshape the encoder itself through fairness-aware optimization.

6.1 Data-centric interventions

6.1.1 Diversify and expand datasets

Systematic underrepresentation of specific speaker demographics, linguistic varieties, and pathological conditions constitutes a primary source of algorithmic bias (Paullada et al., 2021). Addressing this sampling bias requires not merely increasing data volume, but structurally reshaping training distributions to cover the long tail of acoustic variation.

Participatory Data Collection. Involving marginalized communities as collaborators rather than passive subjects improves both the ecological validity of collection protocols and the accuracy of annotation. Co-designing recording tasks with target populations ensures that prompts, pacing, and interfaces match real communication routines. Aging research workshops refined protocols to avoid younger-user defaults (Hu et al., 2025b); Project Euphonia adapted interfaces to dysarthric speakers’ fatigue constraints (MacDonald et al., 2021; Martin et al., 2025); and co-design frameworks empowered low-resource language communities to set culturally grounded collection criteria (Huaman et al., 2026; Emezue et al., 2025). Granting communities authority over annotation is equally important. Standard labeling pipelines often fail for non-standard dialects and low-resource languages because annotators lacking lived context may normalize forms or mistranscribe code-switching (Harris et al., 2024; Martin & Wright, 2023). Community-led curation, from NaijaVoices’ native-speaker transcription conventions (Emezue et al., 2025) to Common Voice’s crowdsourced validation (Ardila et al., 2019), reduces such label noise and improves representativeness for minoritized speakers (Kuhn et al., 2020; Mitra et al., 2016; Reitmaier et al., 2024; Li & Wu, 2024b).

Augment Data. When collection is infeasible, augmentation can populate sparse acoustic regions via synthesis or signal transformation. For clinical speech, TTS-based synthesis generates utterances that preserve speaker and disorder characteristics: Atty-TTS (Wang et al., 2023a) for Spoken Language Understanding (SLU) and TTDS (Leung et al., 2024) for dysarthric ASR. VC extends this to low-resource languages by producing dysarthric variants without target-language clinical data (Li et al., 2025a), and StutterAug (Singh Yadav et al., 2024a) simulates disfluencies to reduce the elevated false positive rates that synthetic speech detectors exhibit on speech-impaired speakers. For accent fairness, cross-lingual VC generates non-native variants (Zhang et al., 2022e; 2023b); combining VC with speed perturbation and SpecAugment can yield measurable bias reduction for non-native-accented ASR (Zhang et al., 2023c). For age-related gaps, VC-based methods convert adult speech into child-like variants or increase inter-child diversity (Zhang et al., 2024b; Zhao et al., 2023; Singh et al., 2021; Shahnawazuddin et al., 2020), while vocal tract length normalization and speed perturbation address broader age and accent disparities (Patel & Scharenborg, 2023). Pitch and formant manipulation address gender data imbalance by augmenting underrepresented f0 ranges (Fucci et al., 2023). These techniques can also operate without demographic labels: López et al. (López et al., 2026) pair spectrum-level augmentation with knowledge distillation to reduce demographic disparity in wake-word detection.

Both strategies have inherent limits. Participatory collection is costly and difficult to scale beyond individual communities, while synthetic augmentation risks introducing generator artifacts that do not reflect the true acoustic distribution of the target group, potentially trading one form of bias for another.

6.1.2 Strategic Resampling and Rebalancing

While expanding datasets is ideal, addressing imbalances in existing corpora often necessitates *strategic resampling*. The standard empirical risk minimization (ERM) objective naturally biases optimization towards majority groups, as their gradients dominate the parameter updates. Resampling techniques interventionally alter the training distribution, typically via *downsampling* majority cohorts or *upsampling* minority ones, to ensure that underrepresented acoustic profiles exert a proportional influence on the learning trajectory.

In the speech domain, this strategy serves as a critical baseline across diverse tasks. For ASR and SV, rebalancing the training distribution, whether by downsampling dominant cohorts or oversampling under-represented ones, has proven effective in mitigating bias across gender (ElGhazaly et al., 2025; Fenu et al., 2021), age (Maison & Estève, 2023), and accent (DHERAM et al., 2022; Garg et al., 2024; Estevez & Ferrer, 2023). By preventing the model from collapsing into majority-class priors, this approach effectively equalizes performance metrics such as False Rejection Rates. Furthermore, in pathological speech detection, rebalancing represents a necessity rather than a simple optimization heuristic. Healthy recordings dominate clinical corpora, and classifiers trained on such skewed data perform poorly on the pathological minority class; undersampling the healthy majority or oversampling the minority substantially recovers this performance (Núñez-Vidal et al., 2024; Lee & Lee, 2023; Zhang et al., 2022b).

However, resampling is not without cost. Aggressive downsampling improves fairness metrics often at the expense of *data efficiency* and potentially global accuracy, as it discards informative samples. Conversely, naive upsampling risks overfitting to the limited diversity of minority examples. These costs also help explain why ASR work rarely downsamples healthy speech, repurposing it instead as an augmentation resource for underrepresented groups.

6.1.3 Data selection and curriculum learning

Strategic resampling corrects quantitative imbalance, but treats all samples within a group as equally informative. **Data selection and curriculum learning** complement resampling by focusing on the utility of individual samples: a hard or underrepresented example provides more valuable gradient information than an easy or prototypical one. Therefore, these algorithms dynamically prioritize samples based on learning difficulty, diversity, or uncertainty.

Error-Driven Prioritization. Approaches inspired by the *Just Train Twice* framework (Liu et al., 2021a) use training difficulty as a proxy for marginalization. In speech, the same principle supports selective preprocessing that is applied only under disparity-inducing conditions, such as DENOASR’s targeted denoising (Rai et al., 2024). It also motivates gap- and divergence-aware data reallocation, including privacy-preserving selection and acquisition policies that prioritize collecting or labeling utterances most likely to close subgroup gaps under limited budgets (Koudounas et al., 2025b;a; 2024a;b). Complementarily, confidence-guided augmentation can inject counterfactual diversity around low-confidence regions, as in CO-VADA (Tsai et al., 2025).

Targeted Subset Selection. When computational or annotation budgets are constrained, selecting the mathematically optimal subset for fairness becomes critical. Algorithms like DITTO (Kothawade et al., 2023) utilize submodular mutual information to curate training batches that maximize acoustic diversity while matching a target distribution.

Fairness-Aware Active Learning. Unlike static filtering, active learning dynamically triggers the acquisition of new labels to address representation gaps. An active-learning approach for accent adaptation prioritizes utterances matching the target dialect by augmenting uncertainty sampling with a cross-entropy relevance criterion (Nallasamy et al., 2012).

6.1.4 Counterfactual Data Augmentation

Counterfactual data augmentation operationalizes causal fairness by synthesizing matched “what-if” speech pairs where a sensitive attribute is altered while task-relevant attributes remain fixed (Kusner et al., 2017). The pairs either enforce consistency regularization, penalizing the model when its predictions change between factual and counterfactual inputs, or simply rebalance the training data with attribute-flipped copies.

In ASR, *Counterfactually Fair ASR* (Sar et al., 2021) flips gender, age, education, and race via adversarial auto-encoder voice conversion and trains with the consistency penalty. Fucci et al. (Fucci et al., 2023) achieve the flip with signal processing alone, shifting f0 and formants across gender-typical ranges; ablations confirm that the cross-gender flip, not generic pitch perturbation, drives the fairness gain. Voice-preserving accent conversion extends the idea, turning native utterances into seven non-native accents, though gains are limited to accents seen during augmentation and vanish for models pretrained on native speech (Klumpp et al., 2023). CO-VADA (Tsai et al., 2025) addresses bias in SER by generating identity-shifted counterfactuals, forcing

the model to unlearn spurious correlations between vocal timbre and emotion labels. However, not every conversion-based augmentation is counterfactual. Since conventional voice conversion preserves the accent of its source speech, Zhang et al. (Zhang et al., 2022e; 2023b) flip speaker identity while the accent stays fixed, diversifying voices within the underrepresented group rather than constructing counterfactuals over the sensitive attribute. All these methods assume the transformation alters only the target attribute; in practice, conversion systems also perturb correlated factors such as prosody and speaking rate, weakening the causal validity of the counterfactuals.

6.2 Feature-level methods

These methods offer a lightweight alternative to full retraining: they can be applied retroactively to any pretrained model, making them especially attractive for large self-supervised systems where end-to-end optimization is prohibitive.

6.2.1 Geometric Subspace Projection

Geometric approaches offer a stable alternative to adversarial training, removing sensitive directions from the embedding space with closed-form linear algebra. Krishnan et al. (Krishnan et al., 2024) identify a “bias subspace” of directions maximally correlated with sensitive attributes in self-supervised models and project embeddings onto its orthogonal complement, removing demographic leakage without retraining or with little loss of linguistic utility. A privacy-motivated variant whitens x-vectors along the sex-discriminant direction, removing most gender information (Noe et al., 2022). The operation itself is native to speaker verification, where nuisance-attribute projection has removed channel and language variability for two decades (Solomonoff et al., 2004; Silnova et al., 2022); fairness redirects it at demographic subspaces. However, such projections remove only linearly encoded bias; nonlinear demographic information may survive.

6.2.2 Feature reweighting

Feature suppression approaches aim to explicitly localize and attenuate bias-encoding dimensions. These methods operate on the hypothesis that sensitive attributes are concentrated in specific spectro-temporal regions or latent feature subsets. By identifying and masking these “biased regions,” models are forced to rely on robust, invariant cues. Realizations of this strategy employ saliency-driven gradient localization to identify input patches correlated with age or accent, thereby redirecting the model’s focus toward intrinsic task features (Wang et al., 2023c; Sameti et al., 2026). The same principle extends to hand-crafted feature sets: Yang et al. (Yang et al., 2024) use statistical tests to rank eGeMAPS features by their gender discriminability and remove the most correlated ones, reducing gender-dependent prediction disparities.

While computationally more stable than adversarial training, these methods face the localization bottleneck. They rely heavily on the fidelity of the attribution method. If the bias regions overlap with content regions, such as when gender cues are entangled with emotion pitch, suppression can lead to significant degradation in task utility.

6.2.3 Model compression

Model compression techniques such as pruning, quantization, and distillation can implicitly regularize away spurious correlations by reducing model capacity, thereby limiting the encoding of demographic cues that drive unfair predictions. Compressed models often correlate less with sensitive attributes (Xu & Hu, 2022), and in speech SSL representations, row pruning provides the most consistent debiasing effect (Lin et al., 2024b). However, indiscriminate compression risks “disproportionate forgetting,” concentrating performance degradation on underrepresented subgroups (Hooker et al., 2019; Proskurina et al., 2023). Strictly speaking, compression modifies encoder weights and thus sits at the boundary of our feature-level definition; we group it here because it is applied post hoc to a trained model without fairness-specific retraining.

Compared with NLP, where iterative nullspace projection (Ravfogel et al., 2020) and closed-form linear erasure (Belrose et al., 2023) have become standard post-hoc debiasing tools, feature-level methods in speech remain few. Speech embeddings are continuous and variable-length, and demographic cues tend to be more

tightly entangled with linguistic content than in text representations, making direct adaptation of these algebraic techniques a nontrivial and largely unexplored problem.

6.3 Training-level algorithms

This is the largest category in the speech fairness literature. We order the methods along a spectrum: from suppressing demographic information in the learned representation, through constraining the loss function for equitable outcomes, to explicitly incorporating demographic knowledge into the model, and finally to adapting pretrained models for underserved subgroups.

6.3.1 Adversarial Disentanglement

Adversarial disentanglement formulates fairness as a minimax optimization problem. The goal is to learn a representation that maximizes utility for the primary task while minimizing the information available to an auxiliary discriminator regarding sensitive attributes. Formally, this is typically achieved via a Gradient Reversal Layer (GRL) (Ganin & Lempitsky, 2014), which encourages the encoder to remove demographic leakage, thereby learning representations that are invariant to group membership (Madras et al., 2018).

Standard Gradient Reversal Framework: The foundational strategy employs a simple GRL to backpropagate negative gradients from a sensitive attribute discriminator. This baseline architecture is widely used to normalize acoustic variability in ASR (Sun et al., 2018; Tripathi et al., 2018; Hsu et al., 2024; Shrivastava et al., 2025; Woszczyk et al., 2020; Das et al., 2021), align representations across speaker groups (e.g., child-adult) (Lahiri et al., 2020), and enforce age-invariance in SV (Qin et al., 2022). In paralinguistic tasks, it serves to strip gender or identity cues from depression detection (Kim et al., 2025b), and SER models (Chien & Lee, 2023; Kasun et al., 2022; Upadhyay et al., 2025; Cai et al., 2021). The same principle extends to music: Gu et al. (Gu et al., 2023) combine adversarial training with conditional alignment to enforce gender invariance in singing voice transcription.

Generative and Reconstruction-Based Architectures: To prevent feature collapse, generative approaches enforce a reconstruction constraint alongside the adversarial objective. For instance, AIPNet (Chen et al., 2020) and DAAE (Gao et al., 2021) utilize generative adversarial pre-training to separate accent- or speaker-invariant features. Similarly, LR-VAE (Huang et al., 2021) and FS-VAE (Huang et al., 2022b) integrate GRL-based adversarial purification with variational autoencoder (VAE)-style reconstruction to suppress sensitive-attribute leakage without discarding essential speech information.

Attention-Guided Adversarial Learning: Standard GRL applies uniform gradient reversal to all frames, which can under-suppress bias cues concentrated in specific regions. Attention-guided variants address this by learning to focus the adversarial signal on bias-salient dimensions. AADIT (Meng et al., 2019) weights deep features by their importance to domain classification, while AANet (Yang et al., 2022a) adds an adaptive attention module after the encoder to selectively remove accent information. Both reduce ASR disparities across recording conditions and accents, respectively.

This family carries two known caveats. Minimax optimization can be unstable and sensitive to the discriminator’s capacity, and the removal is often incomplete, since post-hoc probes can still recover sensitive attributes from adversarially purified representations (Elazar & Goldberg, 2018).

6.3.2 Geometric Alignment via Contrastive Learning

Contrastive debiasing mitigates bias by shaping the embedding geometry to enforce invariance, ensuring that samples sharing the same target label are aligned across diverse demographic groups while attribute-specific variability is suppressed. One paradigm relies on supervised cross-group pairing, where models explicitly align representations of parallel utterances that share task annotations but differ in sensitive attributes like accent or gender (Han et al., 2021; Chien et al., 2024). In contrast, alignment without parallel data regularizes latent clusters or distributions to prevent the geometric isolation of underrepresented subgroups, though group labels may still be required (Koudounas et al., 2024a; Kim et al., 2025a). Advancing to finer granularity, frameworks like DyPCL apply these objectives at the phoneme level, capturing local acoustic nuances critical for clinical or atypical speech conditions (Lee et al., 2025b).

6.3.3 Optimization Objective Modification

This paradigm alters the loss function or gradient update trajectory to constrain the model’s behavior, ensuring equitable performance across groups without necessarily changing the model architecture.

Regularization for Distribution Parity. These approaches augment the standard training objective with a penalty term, formulated as $\mathcal{L}_{total} = \mathcal{L}_{task} + \lambda\mathcal{L}_{fair}$, to constrain the model’s optimization trajectory toward fairness. Depending on the specific task requirements, these constraints target either statistical independence or error rate parity.

To enforce *demographic parity*, researchers penalize the divergence of model predictions; for instance, Kim et al. minimized the Wasserstein distance between demographic subgroups in automated video interviews (Kim et al., 2023b), while Chatteraj et al. utilized a heterogeneity regularizer in public speaking assessment to decouple sensitive attributes from performance ratings (Chatteraj et al., 2020). Conversely, when enforcing *equalized odds* (equal error rates are preferred), approaches like FairM2S integrate equalized odds constraints into a meta-learning framework to penalize gender-dependent disparities in stress detection (Shelke et al., 2025). Similarly, for ASR, Gao et al. proposed a modified Inclusive CTC loss to balance recognition performance across diverse accents (Gao et al., 2022). In general, similar objective-based regularization strategies have been successfully applied to mitigate bias in SV (Sharma, 2024) and SER (Gorrostieta et al., 2019; Lin et al., 2025c; Chou et al., 2024). Going further, minimax-style variants move beyond a fixed criterion and instead penalize the worst-case inter-group gap: ERM-MinMaxGAP (Pang et al., 2026) minimizes the largest male-female loss disparity across languages in SLM-based SER.

Reweighting and Distributionally Robust Optimization (DRO). These methods adjust per-sample contributions to the loss to counteract data imbalance. Static approaches use inverse-frequency weighting, while dynamic variants train an auxiliary bias-capturing model that upweights samples conflicting with majority shortcuts (Lin et al., 2025c), or employ an adversarial network to surface and reweight poorly performing demographic groups (Jin et al., 2022). DRO formalizes this intuition by directly optimizing worst-case subgroup performance (Sagawa et al., 2020); CTC-DRO (Bartelds et al., 2025) further adapts group DRO to CTC-based multilingual ASR; EMO-Debias (Lin et al., 2025c) applies DRO to strictly bound error rates of the most disadvantaged groups in SER. When demographic labels are unavailable, unsupervised cluster-guided methods infer latent acoustic subgroups as proxies, discovering and upweighting emerging worst-case clusters in ASR, IC, and SER (Veliche & Fung, 2023; Koudounas et al., 2024a; Lin et al., 2025b).

Information-theoretic constraint. Information-theoretic approaches reduce bias by explicitly formulating fairness as an optimization of the Information Bottleneck (Colombo et al., 2023). The objective is to learn a representation that maximizes mutual information (MI) with task-relevant information while minimizing MI with sensitive cues. A recent application is WavShape (Baser et al., 2025), which uses MI estimators (Belghazi et al., 2018) to suppress sensitive information and preserve linguistic content, yielding privacy-aware and demographically invariant speech representations. However, MI is difficult to estimate reliably in high-dimensional continuous embeddings; variational estimators can be sensitive to critic capacity, batch size, and optimization, and may require careful tuning to avoid unstable training (McAllester & Stratos, 2020).

6.3.4 Explicit Attribute Modeling

Rather than ignoring sensitive attributes, this paradigm explicitly incorporates them into the model to account for systematic variations in speech production.

Bias-Aware Conditioning. These methods supply demographic or speaker identity signals as additional model inputs so that the network can account for systematic variation across groups. At the *group level*, conditioning signals include discovered speaker-cohort membership for ASR (DHERAM et al., 2022), parallel group-adapted encoders with score fusion for SV (Shen et al., 2022), and one-hot-encoded age, gender, and education attributes fused with acoustic embeddings via adaptive gating for cognitive screening (Cog.) (Zolnoori et al., 2025). In speech translation, Gaido et al. (Gaido et al., 2023) inject speaker gender metadata and gender-specific external language models to prevent masculine-default outputs, while Bansal et al. (Bansal et al., 2025) eliminate the explicit gender input by fine-tuning on LLM-corrected references so the model infers gender from acoustic cues alone. For low-resource dialects, Tsai et al. (Peng et al., 2026) condition ASR

on dialect embeddings to bridge performance gaps across Hakka varieties that standard models underserve. At the *individual level*, Triantafyllopoulos and Schuller (Triantafyllopoulos & Schuller, 2024) use attention over enrollment utterances to personalize emotion recognition, capturing individual idiosyncrasies that coarse group labels miss.

Multi-task Learning (MTL) and Auxiliary Learning. While adversarial disentanglement (§6.3.1) strips demographic cues to achieve invariance, MTL takes the complementary view that explicitly modeling group variation as a structured factor better serves tasks where demographic differences entail legitimate acoustic variation. By jointly optimizing the primary objective alongside auxiliary heads that predict group-related attributes, the shared encoder is forced to capture subgroup-specific patterns rather than overfitting to the dominant demographic (Wang & Sun, 2024; Montalvo et al., 2020; Peri et al., 2023). In accented ASR, auxiliary accent classification heads compel the model to distinguish valid pronunciation variants from recognition errors, boosting accuracy for non-native speakers (Zhang et al., 2021; Xiao et al., 2025) and dialects (Zhao et al., 2019; Imaizumi et al., 2022; Dan et al., 2022). The same principle extends to low-resource language transfer (Yolwas & Meng, 2023) and multilingual TTS, where a language identification auxiliary preserves prosodic structures for diverse inputs (Zhang et al., 2022a).

Interpretable Feature Routing. Rather than soft conditioning or auxiliary losses, Fair-Gate (Qu et al., 2026) introduces learnable routing masks that split intermediate features into identity and sex branches, combined with risk equalization across sex groups to close SV performance gaps.

Most methods in this subsection presuppose demographic annotations at training time, whether as conditioning signals, auxiliary labels, or routing targets; the individual-level enrollment scheme above (Triantafyllopoulos & Schuller, 2024) is the rare exception, replacing group labels with the target speaker’s own recordings. For the rest, missing or privacy-restricted annotations block training outright, while noisy annotations degrade the learned group models.

6.3.5 Targeted Adaptation Strategies

Standard training on pooled data often overfits to the majority distribution. Rather than retraining from scratch, targeted adaptation strategies selectively update a pretrained model, whether through partial fine-tuning, parameter-efficient modules, or dynamic routing, to recover performance on underserved subgroups. In pronunciation assessment, for example, Nijat et al. (Nijat et al., 2026) show that the native-norm ground truth used by standard mispronunciation detectors is itself a fairness risk. Selectively fine-tuning only the phonetic module on target population (L2) speech shifts the reference to a listener-adapted norm, improving both detection accuracy and cross-L1 parity.

Parameter-Efficient Fine-Tuning (PEFT). PEFT learns small, subgroup-specific parameter updates on top of a shared backbone, capturing missing acoustic factors without overwriting general competence. Residual adapters that update less than 0.5% of parameters were first shown to match full fine-tuning for atypical and accented speakers (Tomanek et al., 2021). Subsequent applications span child speech (Liu et al., 2024), speakers with autism spectrum disorder (Park & Lee, 2025), per-speaker adapter fusion for dysarthric ASR (Qi & Van hamme, 2023), and inclusive Low-Rank Adaptation (LoRA) for low-resource groups (Acharya et al., 2025).

Beyond adapting to a single target group, PEFT can embed fairness objectives directly: Swain et al. (Swain et al., 2026) couple fairness-prompted adapters with Spectral Decoupling, group DRO, and invariant risk minimization to equalize WER across 26 L2 accent groups, while benchmark evidence from SER shows that LoRA improves demographic fairness alongside accuracy relative to other PEFT variants (Feng & Narayanan, 2023). A known risk is that naive subgroup finetuning can overfit to a handful of speakers, creating new bias; factorized deficiency adapters mitigate this by improving within-subgroup generalization (Hu et al., 2024a).

Mixture-of-Experts (MoE) Routing. Instead of assigning a fixed adapter per subgroup, MoE learns a router that dynamically selects or combines experts per input, allowing subgroup structure to emerge from the data. Hu et al. (Hu et al., 2025a) initialize experts with severity- and gender-conditioned adapters and learn on-the-fly routing over HuBERT/WavLM for dysarthric ASR. MOPSA (Deng et al., 2025) clusters

speaker-adaptive prompts into prompt-experts via k -means and routes them at both the encoder and decoder levels of Whisper for elderly speech.

Both paradigms assume that subgroup boundaries are either known a priori (PEFT) or at least partially discoverable from acoustic clustering (MoE). When subgroup definitions are ambiguous or shift across deployment contexts, the resulting specialization may not transfer.

6.4 Inference and post-processing

Unlike training-centric approaches that require computationally expensive model retraining, inference-time interventions operate during the deployment phase. These methods mitigate bias by adapting the input signal, guiding the decoding process, or calibrating the final output scores. They offer a "plug-and-play" mechanism for fairness, particularly valuable when accessing model parameters is infeasible.

6.4.1 Test-time Input Adaptation

While voice conversion is used at training time to construct counterfactual pairs (§6.1.4), it can also serve as a real-time preprocessing step that neutralizes demographic cues before the signal reaches the downstream model. Rizhinashvili et al. (Rizhinashvili et al., 2022) apply voice transformation to strip gender-specific characteristics from speech, Alhumud et al. (Alhumud et al., 2024) use VC to transform accented speech toward a native accent, and Székely et al. (Székely et al., 2023) synthesize gender-ambiguous voices whose prosodic cues are neutralized. These methods work because deployed models are most accurate on the majority-group speech that dominates their training data; converting underrepresented inputs toward majority-group characteristics can therefore recover performance without new data or retraining. The same mechanism is also this family’s normative weakness. Adapting the user rather than the model enacts the acoustic assimilation critiqued in §3.2, trading representational harm for recognition accuracy, and is therefore best treated as a pragmatic stopgap when the model itself cannot be modified.

6.4.2 Test-time Representation Intervention

Rather than adapting inputs or parameters, these methods directly manipulate frozen model representations at inference time to reduce demographic disparities. Nachesa and Niculae (Nachesa & Niculae, 2025) augment Whisper (Radford et al., 2023) with a token-level k -nearest-neighbor datastore and show through explicit demographic analysis that retrieval-augmented decoding yields differentiated WER improvements across gender, accent, and age groups. In a different vein, activation steering injects learned directional vectors into targeted hidden layers without modifying weights. Sun et al. (Sun et al., 2026) extract accent-induced mean-shift directions and inject them into encoder layers, achieving consistent WER reductions across eight accents. Yang and Hansen (Yang & Hansen, 2026) extend this to zero-shot TTS by steering activations to produce accent-neutralized synthesis.

6.4.3 Test-time Parameter Adaptation

These methods estimate a small set of subgroup-dependent parameters from test data at inference time, specializing the model for underrepresented demographics without full retraining. On-the-fly feature-based Learning Hidden Unit Contributions (f-LHUC) (Geng et al., 2023; 2025) derives speaker-dependent transforms and variance-regularized spectral basis embeddings from minimal dysarthric or elderly test audio, reducing WER without gradient-based optimization. For children’s speech, Shi et al. (Shi et al., 2025) show that applying prior Test-time Adaptation (Lin et al., 2022; Kim et al., 2023a) methods can meaningfully narrow the child–adult performance gap.

6.4.4 Decoding and Rescoring Strategies

Modifying the search or scoring process at inference time can correct demographic disparities without retraining. Cross-utterance rescoring (Tankasala et al., 2023) builds an acoustic similarity graph across utterances from multiple speakers and applies label propagation so that high-confidence hypotheses from well-served accents guide recognition of underrepresented ones, explicitly mitigating majoritarian bias.

Complementarily, Southwell et al. (Southwell et al., 2024) add a speed-aware rescoring term to beam search that penalizes hypotheses deviating from expected child speaking rates, providing an additional decoding-time gain on top of model fine-tuning for child ASR.

6.4.5 LLM-Driven Correction

Leveraging the semantic reasoning of frozen LLMs, recent frameworks mitigate bias through inference-time post-processing rather than expensive model retraining. In medical domains, context-aware prompting and Chain-of-Thought reasoning have proven effective in mapping accent-induced phonetic errors back to valid terminology for non-native speakers (Fatapour et al., 2025; Adedeji et al., 2025). Parallel efforts extend this zero-shot capability to dysarthric speech, utilizing multimodal LLMs to decode disordered patterns and bridge the digital divide for users with disabilities (Alsayegh & Masood, 2025).

6.4.6 Threshold calibration

Threshold calibration adjusts decision boundaries or score distributions after training to equalize error rates across demographic groups (Hardt et al., 2016). In SV, a single global threshold disproportionately harms underrepresented groups: calibration performance degrades far more severely than discrimination performance on out-of-domain accents (Estevez & Ferrer, 2023), and the choice of operating point further reshapes the fairness-security-usability trade-off across demographic groups (Fenu et al., 2020a). Bias audits confirm that such disparities persist throughout the decision pipeline (Hutiri & Ding, 2022). Crucially, these disparities can be mitigated without modifying the underlying model. Jain and Wang (Jain & Wang, 2021) fit gender- and age-specific decision thresholds on held-out scores from a frozen SV system, reducing false rejection rates for disadvantaged subgroups. Estevez et al. (Estevez et al., 2025) similarly apply group-specific posterior calibration to a fixed voice disorder detector, correcting age- and gender-dependent miscalibration that global metrics fail to reveal. Because only the post-hoc score-to-decision mapping changes, both methods are applicable at inference time to any pre-trained system.

6.4.7 Prompting and In-Context Adaptation

With the advent of Large Audio Language Model (LALM), fairness can be elicited through context manipulation rather than parameter updates. Peng et al. (Peng et al., 2023) utilize discrete prompting, demonstrating that injecting task-specific tokens effectively activates dormant capabilities for code-switching and minority linguistic patterns. Advancing to demonstration-based learning, Roll et al. (Roll et al., 2025) and Wang et al. (Wang et al., 2024a) show that prepending interleaved audio-text or speech-only exemplars enables robust zero-shot adaptation for low-resource English and Chinese dialects, respectively. To address severe data scarcity, Hsu et al. introduces *SMILE* to meta-learn adaptation cues from unlabeled audio, while Li and Niehues leverage *Multimodal In-Context Learning (MICL)* to facilitate cross-lingual transfer for unseen languages by synergizing speech and text contexts (Li & Niehues, 2026). Collectively, these mechanisms allow frozen models to dynamically align with diverse demographic subgroups during inference, offering a scalable and efficient alternative to computationally expensive fine-tuning. These ideas also extend to generation: CLARITY (Poon et al., 2026) combines LLM-guided text localization with retrieval-augmented accent prompting to mitigate accent bias in zero-shot TTS.

Table 5: Summary of mitigation strategies for speech fairness. **Paradigm** follows §3: Rob. = robustness; Rep. = representation; Gov. = governance. **Bias Source** follows §5: *Data* = data-related bias (§5.1); *Entangl.* = feature entanglement (§5.2.1); *Capacity* = capacity limitations (§5.2.1); *Decoding* = decoding assumptions (§5.2.2); *Deploy.* = deployment mismatch (§5.3). **Supervision**: signal required beyond task labels. **Model Req.**: None = data-only; Post-hoc = applied to a trained model without retraining; Loss = loss change; Arch. = architecture change; Infer. = inference-time access.

Strategy	Paradigm	Bias Source	Supervision	Model Req.	Tasks	Rep. Work
<i>A. Data-centric interventions</i>						
Diversify & expand	Rob., Gov.	Data	Task labels	None	ASR, SLU, SSD	(Wang et al., 2023a; Leung et al., 2024; Zhang et al., 2022e)

Strategy	Paradigm	Bias Source	Supervision	Model Req.	Tasks	Rep. Work
Strategic resampling	Rob.	Data	Group labels	None	ASR, SV, SER	(DHERAM et al., 2022; Estevez & Ferrer, 2023; ElGhazaly et al., 2025)
Data selection & curriculum	Rob.	Data	Loss signals	None	ASR, SER, IC	(Koudounas et al., 2025b; Kothawade et al., 2023)
Counterfactual augmentation	Rob., Rep.	Data / Entangl.	Paired data	None	ASR, SER	(Sar et al., 2021; Fucci et al., 2023; Tsai et al., 2025)
<i>B. Feature-level methods</i>						
Geometric projection	Rob., Rep.	Entangl.	Group labels	Post-hoc	ASR, SV	(Krishnan et al., 2024)
Feature reweighting	Rob., Rep.	Entangl.	Attribution maps	Post-hoc	ASR, clinical	(Wang et al., 2023c; Sameti et al., 2026)
Model compression	Rob.	Entangl.	Task labels only	Post-hoc	ASR	(Xu & Hu, 2022; Lin et al., 2024b)
<i>C. Training-level algorithms</i>						
Adversarial disentanglement	Rob., Rep.	Entangl.	Group labels	Arch.	ASR, SV, SER, clinical	(Sun et al., 2018; Qin et al., 2022; Kim et al., 2025b)
Contrastive alignment	Rob., Rep.	Entangl.	Group labels / unsup.	Loss	ASR, SER, clinical	(Han et al., 2021; Kim et al., 2025a; Lee et al., 2025b)
Objective modification	Rob.	Data / Entangl.	Group labels / unsup.	Loss	ASR, SV, SER, IC	(Gao et al., 2022; Bartelds et al., 2025; Shelke et al., 2025)
Explicit attribute modeling	Rob., Rep.	Capacity	Group labels	Arch.	ASR, SV, ST, SER, Cog., TTS	(DHERAM et al., 2022; Shen et al., 2022; Qu et al., 2026)
Targeted adaptation	Rob.	Capacity	Group labels / clusters	Arch.	ASR	(Swain et al., 2026; Hu et al., 2025a)
<i>D. Inference and post-processing</i>						
Input adaptation	Rob.	Entangl. / Deploy.	None	Infer.	SV, ASR	(Rizhinashvili et al., 2022; Alhumud et al., 2024)
Representation intervention	Rob., Rep.	Capacity / Entangl.	Datastore / directions	Infer.	ASR, TTS	(Nachesa & Niculae, 2025; Sun et al., 2026)
Parameter adaptation	Rob.	Deploy.	Unlabeled test data	Infer.	ASR	(Geng et al., 2023; Shi et al., 2025)
Decoding & rescoring	Rob.	Decoding	Task heuristics	Infer.	ASR	(Tankasala et al., 2023; Southwell et al., 2024)
LLM-driven correction	Rob.	Data / Entangl.	None (ext. LLM)	Infer.	ASR	(Fatapour et al., 2025; Alsayegh & Masood, 2025)
Threshold calibration	Rob.	Decoding	Group labels	Infer.	SV, clinical	(Jain & Wang, 2021; Estevez et al., 2025)
Prompting & ICL	Rob.	Deploy.	Few-shot exemplars	Infer.	ASR, TTS	(Peng et al., 2023; Poon et al., 2026)

Summary. Table 5 maps each strategy to the paradigm it serves (§3), the bias source it targets (§5), the supervision it requires, the level of model access it assumes, and the tasks on which it has been validated. We subdivide model-related bias into feature entanglement, capacity limitations, and decoding assumptions because each gives rise to a distinct mitigation family; data-related and deployment bias are kept at the coarser level because their sub-sources do not differentiate mitigation strategies in the same way.

Three patterns emerge. First, all 19 strategies serve robustness, but only seven also address representational fairness, and only one touches governance, reflecting the field’s heavy tilt toward performance parity. Second,

feature entanglement attracts the widest range of solutions across all four pipeline stages, while deployment mismatch remains under-addressed with only inference-time methods available. Third, no single stage is sufficient: data interventions cannot prevent biases introduced by the model itself, training-level algorithms frequently require demographic annotations that may be unavailable or privacy-sensitive, and inference-time methods can only partially compensate for a frozen model’s learned biases. These complementary gaps motivate the open challenges discussed next.

7 Challenge and Future Work

We identify seven open directions, ordered by their dependencies.

Standardized reporting (§7.1) and corpus governance (§7.2) provide the empirical and legal prerequisites for subsequent technical scaling. Without the high-fidelity metadata and non-extractive pipelines they establish, the field cannot reliably navigate the Pareto frontier (§7.3) or leverage controllable data synthesis (§7.4) without compromising user trust. These data-centric foundations, in turn, facilitate surgical model-level innovations: advancements in mechanistic interpretability (§7.5) and test-time adaptation (§7.6) depend on the robust, well-documented baselines established in preceding layers to enable precise, real-time interventions. Finally, fairness for generative AI (§7.7) represents a holistic frontier that integrates these dependencies, necessitating a systemic rethinking of accountability across the entire sociotechnical pipeline.

7.1 Standardized Reporting Protocols

Rigorous comparison across multidimensional fairness benchmarks depends critically on standardized reporting protocols. Without sufficient contextualization, fairness claims remain difficult to interpret and can be actively misleading (Blodgett et al., 2020). To address these ambiguities, the field necessitates the establishment of a systematic reporting framework (Richter et al., 2024) that directly targets the data-related bias sources identified in §5.1, encompassing three categories of speech-centric metadata.

First, **recording context** must be documented to address channel bias (§5.1-2). This includes channel characteristics and acoustic environments, such as whether speech is captured through telephony systems or studio-quality microphones (Gong et al., 2023). Second, **sociolinguistic profiles** are required to mitigate demographic underrepresentation and linguistic disparities (§5.1-1). These profiles should capture essential variables, including language variety and speaker proficiency, which substantially influence system behavior (Richter et al., 2024). Third, **annotator demographics** should be disclosed to surface potential annotation subjectivity (§5.1-3) and cultural alignment bias in the construction of ground truth labels (Sachdeva et al., 2022). Standardizing these metadata forms enables researchers to distinguish genuine algorithmic improvements from dataset-specific artifacts, ensuring that **robustness** is evaluated with appropriate sociotechnical nuance.

7.2 Ethical Corpus Governance

Building on the governance framework established in §3.3, which identified **data sovereignty** as a foundational governance requirement, future research must shift from conceptual advocacy toward technical and institutional implementation. While achieving robust performance requires diverse data, scaling collection without formal governance risks perpetuating the extractive practices and procedural unfairness identified in our systemic analysis. Addressing this imbalance requires operationalizing non-extractive data pipelines grounded in frameworks such as data trusts and the CARE principles (Carroll et al., 2020).

Unlike static open-source datasets, data trusts function as institutional intermediaries that enable marginalized communities to retain ongoing consent and meaningful control over how their biometric data is used (Coudry & Mejias, 2019). To balance such governance with the need for reproducible benchmarks, the field should prioritize the development of **federated evaluation protocols** (Zhang et al., 2023a), where models are audited within community-held environments rather than requiring centralized data aggregation.

Finally, governance must extend to equitable benefit sharing through granular licensing and structural reciprocity. Rather than relying on generic open-source licenses, the field should adopt **community-specific**

licenses that incorporate veto rights over high-stakes commercial applications or the use of Traditional Knowledge Labels (Christen, 2015) to assert ongoing authority over data derivatives. Furthermore, economic value redistribution can be operationalized through royalty-based revenue models (Kukutai & Taylor, 2016) or mandatory reinvestment of licensing fees into community-led language revitalization infrastructure, ensuring that the progress of Speech AI provides direct, tangible returns to the linguistic groups involved.

7.3 Scalable Multi-Objective Evaluation

Fairness and utility trade off along several axes at once (Selbst et al., 2019a), including competing fairness criteria, task performance, and operational cost such as inference latency. Because most evaluations still report a single criterion in isolation, they hide the *Pareto frontier* on which fairness-utility configurations sit (Martínez et al., 2020), motivating *multidimensional benchmarks* that measure several fairness criteria alongside utility and cost (Lin et al., 2025c).

However, navigating this Pareto frontier is inherently search-intensive: identifying optimal fairness-utility configurations requires evaluating a large number of candidate solutions under multiple criteria simultaneously. Standard human-centric metrics such as Δ_{MOS} (§4), as well as subjective listening tests used to assess perceived bias and offensiveness across demographic groups (Satish et al., 2026b), cannot support this throughput because they rely on diverse human evaluator panels that are prohibitively expensive to convene at scale. *Scalable proxies* for human judgment are therefore a functional precondition for tractable multi-objective search. Promising directions include reference-free fairness estimation models and LLM-based evaluators capable of approximating perceptual metrics directly from model outputs (Liu et al., 2023), which would enable the iterative, gradient-based, or evolutionary searches required to discover and maintain optimal fairness-utility balances in complex speech architectures.

7.4 Controllable Data Synthesis

The logistical burden and ethical risks of extractive data collection, particularly from marginalized communities, motivate a shift toward *controllable synthetic data generation* (Chen et al., 2026b). Disentangled representation learning can isolate linguistic content from speaker attributes such as accent, timbre, or pathological markers (Hsu et al., 2017), enabling the synthesis of diverse, high-fidelity speech samples (e.g., mapping dysarthric features onto standard corpora) without compromising the privacy of vulnerable groups.

Such generative frameworks directly address the structural tension where strict data protection laws often starve fairness audits of the very demographic metadata they require (Veale & Binns, 2017). By training models on these de-identified synthetic variations, we move toward a *privacy-preserving fairness* regime that does not rely on identifiable real-world individuals to mitigate bias (Atzori et al., 2025). However, the success of this path hinges on navigating the *fidelity-fairness trade-off*: ensuring that synthetic augmentation does not oversimplify the non-linear acoustic nuances of marginalized speech, which are often lost in current generative approximations.

7.5 Mechanistic Interpretability for Bias

Most existing bias mitigation strategies operate at a coarse level, providing limited insight into how bias is encoded within neural architectures. An important emerging direction is the application of *mechanistic interpretability* to speech models (Plantinga et al., 2025). Techniques such as *causal tracing* (Meng et al., 2022a) and *activation patching* (Facchiano et al., 2025) can localize specific neurons or attention heads that encode stereotypical associations, such as correlations between pitch and perceived social intent (Pasad et al., 2021).

However, applying these methods to speech requires addressing the continuous nature of acoustic signals and the multimodal complexity of modern SLMs. Unlike discrete text tokens, tracing bias in speech must account for temporal dynamics and disentangle whether discriminatory patterns reside in the acoustic encoder or the linguistic decoder. Once localized, these components can be selectively modified through *model surgery* or *rank-one editing* (Meng et al., 2022a), enabling the removal of bias while preserving the model’s general capabilities without exhaustive retraining.

7.6 Test-Time Fairness Adaptation

Deployment mismatch (§5.3) shows that a model trained on a fixed distribution cannot be assumed to generalize across shifting acoustic conditions, speaking styles, and user demographics (Liu et al., 2021b). Test-time adaptation (TTA) offers an alternative to costly retraining, but improving a model’s overall accuracy in a new setting is not the same as improving it fairly for every group of speakers.

Current TTA methods such as TENT (Wang et al., 2021) minimize prediction entropy to raise aggregate accuracy under shift, but these group-agnostic updates can widen the gap for marginalized subgroups rather than close it. Closing this gap calls for test-time objectives that account for subgroup performance, not aggregate accuracy alone. Because demographic labels are often unavailable at inference, promising routes include unsupervised fairness proxies and per-speaker personalization that improve accuracy without sacrificing equity across groups.

7.7 Fairness in the Era of Large Audio Models and Generative AI

Preceding sections have characterized fairness risks in generative speech systems, from representational stereotyping in TTS (§3.2) to content-level bias in SLMs (§4.6). Current frameworks, however, largely evaluate these risks in isolation: on single outputs, in single turns, and after training is complete. Closing these gaps requires research along three directions.

Alignment Bias and Cultural Homogenization: § 3.2 and 5.1 documented how training data and model outputs can privilege standardized linguistic norms. Post-training optimization may introduce an additional homogenization mechanism: supervised fine-tuning teaches models to respond in a narrow range of speaking styles, while preference optimization via RLHF (Ouyang et al., 2022) or DPO (Rafailov et al., 2023) further reinforces whichever style human annotators rate most favorably. In text-based LLMs, this effect is empirically established: alignment has been shown to significantly reduce output diversity (Kirk et al., 2024) and concentrate responses into a small number of semantic clusters (Liu, 2026). Because annotator judgments of speech quality are difficult to disentangle from preferences for prestige varieties, the same optimization pressure may systematically penalize dialectal prosody and code-switching in SLMs, even when pre-training data includes such variation. Whether and to what extent this *cultural homogenization* occurs in current speech systems remains an open empirical question: existing speech alignment work (Zhang et al., 2024a) evaluates quality but does not measure acoustic or linguistic diversity before and after optimization. Establishing this evidence base is a prerequisite for designing effective interventions.

Distributional Fairness in Generative Speech. Generative SLMs define a distribution over possible outputs, and this distribution may contain biased modes whose manifestation depends on the interaction of multiple input factors (Chen et al., 2026a). In text LLMs, decoding parameters such as temperature and sampling strategy have been shown to modulate absolute bias scores (Das & Balke, 2022), and models that appear fair under standard metrics can exhibit systematic bias in their confidence distributions (Wang et al., 2025). These findings suggest that sample-level fairness audits are insufficient for generative systems. Evaluation must move to *distributional* analysis, characterizing whether biased outputs concentrate in the distribution’s tails and whether their probability mass shifts across demographic inputs. In speech, the continuous acoustic output space makes such characterization harder than in discrete text generation.

(c) Multi-turn and Agentic Fairness. In spoken dialogue, acoustic cues such as accent and perceived gender are present in every turn, continuously exposing the model to speaker identity signals that text-based interactions do not carry. Whether this repeated exposure causes demographic-dependent behavior to intensify across turns remains an open empirical question: while multi-turn capability degradation in LLMs is documented (Laban et al., 2025), its interaction with fairness metrics has not been studied for spoken conversations. Future work should develop turn-level fairness tracking methods that measure whether disparities in response quality, verbosity, or tone widen as conversations progress.

Separately, as SLMs are deployed as goal-directed agents that take real-world actions such as scheduling or customer service, fairness must extend beyond conversational content to task outcomes. Whether completion rates, response latency, and error recovery differ systematically across speaker demographics remains a critical

open question (Li et al., 2025b) whose answer will determine whether agentic speech systems amplify or mitigate existing inequities.

8 Conclusion

This survey has organized the fragmented landscape of speech fairness research into a unified framework spanning formal definitions, evaluation metrics, bias sources, and mitigation strategies. Three insights emerge from viewing these layers together.

First, speech fairness is not a single problem but three coexisting paradigms (Robustness, Representation, and Governance), each diagnosing unfairness at a different level. A system that achieves performance parity across demographics may still perpetuate social stereotypes in its outputs or operate within extractive data practices that deny communities agency over their linguistic resources. Recognizing this multiplicity is a prerequisite for comprehensive auditing.

Second, there is a persistent gap between what we can measure and what we can fix. The fairness definitions formalized in this survey give rise to six mathematical families of evaluation metrics, yet many metrics lack corresponding mitigation strategies, and popular interventions such as resampling address robustness but leave representational and governance concerns untouched. The definition-to-family mapping (Table 3) makes these gaps explicit and can guide future work toward currently unaddressed dimensions.

Third, speech presents unique challenges that distinguish it from fairness work in text or vision. Sensitive attributes are acoustically entangled with linguistic content, making them difficult to isolate without degrading utility. Many speech targets are inherently perceptual, so fairness requires scrutinizing annotation protocols and evaluator demographics, not just model behavior.

Section 7 details the open challenges and research directions that follow from these observations. We hope this survey helps the community move from isolated, task-specific audits toward reproducible, diagnosis-driven fairness practice across the speech pipeline.

Limitations

The underlying evidence is concentrated in high-resource languages. The studies, benchmarks, and demographic annotations we build on are dominated by English and a small set of other high-resource languages, because that is where speech-fairness research currently exists. Our conclusions therefore describe disparities as they have been measured in high-resource settings, and the behavior of speech systems for the world’s low-resource and under-documented languages remains under-evidenced, not because it matters less but because the primary studies have not yet been done. This gap is one of the representational inequities the survey identifies, and it bounds the generality of every cross-lingual claim we make.

Our analysis inherits the identity axes that the primary literature chose to measure. A survey of disparities can only be as fine-grained as the studies it synthesizes. Much of the speech-fairness literature records gender as binary, race or first-language background in coarse groupings, and omits attributes such as age, disability, and socioeconomic status, in part because collecting and labeling them raises the very privacy and consent tensions we discuss. Our diagnosis therefore reflects the axes of identity that existing corpora make visible, not the full space of speakers who can be disadvantaged. When we report a specific effect, such as the annotator-gender gap in emotion labels, we are simultaneously reporting the boundary of what current annotations allow anyone to observe.

Our framework is one principled organization, not the only possible one. The three paradigms (Robustness, Representation, and Governance) and the seven definitions are an analytical construction, chosen because they make one specific gap visible: the mismatch between what current metrics measure and what mitigations actually fix. The categories are deliberately not mutually exclusive. Individual studies often span paradigms, and some metrics map to more than one definition. We use these boundaries as a reasoning tool rather than a partition of the field, and a different decomposition could foreground tensions that ours keeps in the background.

The generative frontier moves faster than any survey. The portions of this survey concerned with speech-language models and generative audio describe a rapidly moving target, where both the systems and the fairness behaviors they exhibit change on the timescale of months. We wrote the framework to outlast specific systems: the definitions, the three paradigms, and the pipeline-level diagnosis are intended to remain applicable as models change. Specific empirical findings about today’s speech-language models, however, should be read as a snapshot at the time of writing rather than a settled account.

Acknowledgments

This work was supported by the Ministry of Education (MOE) of Taiwan under the project Taiwan Centers of Excellence in Artificial Intelligence, through the NTU Artificial Intelligence Center of Research Excellence (NTU AI-CoRE). This work was supported in part by the National Science and Technology Council (NSTC), Taiwan, under Grant No. 114-2917-I-564-030 (to Huang-Cheng Chou).

References

- Priyobroto Acharya, Soham Chaudhuri, Sayan Das, Dipanjan Saha, and Dipankar Das. JUNLP@LT-EDI-2025: Efficient Low-Rank Adaptation of Whisper for Inclusive Tamil Speech Recognition Targeting Vulnerable Populations. In *Proceedings of the 5th Conference on Language, Data and Knowledge: Fifth Workshop on Language Technology for Equality, Diversity, Inclusion*, pp. 17–25. Unior Press, sep 2025. ISBN 978-88-6719-334-9. URL <https://aclanthology.org/2025.ltedi-1.4/>.
- Angus Addlesee and Arash Eshghi. You have interrupted me again!: making voice assistants more dementia-friendly with incremental clarification. *Frontiers in Dementia*, 3:1343052, 2024. ISSN 2813-3919. doi: 10.3389/frdem.2024.1343052. URL <https://doi.org/10.3389/frdem.2024.1343052>.
- Ayo Adedeji, Mardhiyah Sanni, Emmanuel Ayodele, Sarita Joshi, and Tobi Olatunji. The Multicultural Medical Assistant: Can LLMs Improve Medical ASR Errors Across Borders?, 2025. URL <https://arxiv.org/abs/2501.15310>.
- Zhiqi Ai, Meixuan Bao, Zhiyong Chen, Zhi Yang, Xinnuo Li, and Shugong Xu. VoxAging: Continuously Tracking Speaker Aging with a Large-Scale Longitudinal Dataset in English and Mandarin. In *Interspeech 2025*, pp. 3648–3652. ISCA, 2025. doi: 10.21437/interspeech.2025-57. URL <https://doi.org/10.21437/interspeech.2025-57>.
- Anas Mohammed Alhumud, Muhammad Al-Qurishi, Yasser Omar Alomar, Ali Alzahrani, and Riad Souissi. Improving Automated Speech Recognition Using Retrieval-Based Voice Conversion. In *Tiny Papers @ ICLR*, 2024. URL <https://openreview.net/forum?id=OMBFB6pU6c>.
- Ranya Aloufi, Srishti Gupta, Soumya Shaw, Battista Biggio, and Lea Schönherr. Evaluation of Audio Language Models for Fairness, Safety, and Security, 2026. URL <https://arxiv.org/abs/2603.13262>.
- Ali Alsayegh and Tariq Masood. Zero-Shot Recognition of Dysarthric Speech Using Commercial Automatic Speech Recognition and Multimodal Large Language Models, 2025. URL <https://doi.org/10.48550/arxiv.2512.17474>.
- Kemal Altwlkany, Amar Kuric, and Emanuel Lacic. On the Language and Gender Biases in PSTN, VoIP and Neural Audio Codecs. In *Interspeech 2025*, pp. 1348–1352. ISCA, 2025. doi: 10.21437/interspeech.2025-481. URL <https://doi.org/10.21437/interspeech.2025-481>.
- Frédéric Aman, Véronique Aubergé, and Michel Vacher. Influence of Expressive Speech on ASR Performances: Application to Elderly Assistance in Smart Home. In *Lecture Notes in Computer Science*, pp. 522–530. Springer International Publishing, 2016. ISBN 9783319455099, 9783319455105. doi: 10.1007/978-3-319-45510-5_60. URL https://doi.org/10.1007/978-3-319-45510-5_60.
- Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common Voice: A Massively-Multilingual Speech Corpus, 2019. URL <https://doi.org/10.48550/arxiv.1912.06670>.

-
- Siddhant Arora, Kai-Wei Chang, Chung-Ming Chien, Yifan Peng, Haibin Wu, Yossi Adi, Emmanuel Dupoux, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. On The Landscape of Spoken Language Models: A Comprehensive Survey, 2025. ISSN 2835-8856. URL <https://doi.org/10.48550/arxiv.2504.08528>.
- Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 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*, pp. 21318–21340. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.1188. URL <https://doi.org/10.18653/v1/2024.emnlp-main.1188>.
- Andrea Atzori, Pietro Cosseddu, Gianni Fenu, and Mirko Marras. The impact of balancing real and synthetic data on accuracy and fairness in face recognition. In *Lecture Notes in Computer Science*, pp. 284–302. Springer Nature Switzerland, 2025. ISBN 9783031919060, 9783031919077. doi: 10.1007/978-3-031-91907-7_17. URL https://doi.org/10.1007/978-3-031-91907-7_17.
- Massa Baali, Abdulhamid Aldoobi, Hira Dhamyal, Rita Singh, and Bhiksha Raj. Pdaf: A phonetic debiasing attention framework for speaker verification. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1209–1216. IEEE, 2024. doi: 10.1109/slt61566.2024.10832294. URL <https://doi.org/10.1109/slt61566.2024.10832294>.
- Massa Baali, Sarthak Bisht, Francisco Teixeira, Kateryna Shapovalenko, Rita Singh, and Bhiksha Raj. SVeritas: Benchmark for Robust Speaker Verification under Diverse Conditions. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9714–9731. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.516. URL <https://doi.org/10.18653/v1/2025.findings-emnlp.516>.
- Rajas Bansal. A survey on bias and fairness in natural language processing, 2022. URL <https://doi.org/10.48550/arxiv.2204.09591>.
- Shubham Bansal, Vikas Joshi, Harveen Chadha, Rupeshkumar Mehta, and Jinyu Li. Addressing speaker gender bias in large scale speech translation systems, 2025. URL <https://arxiv.org/abs/2501.05989>.
- Solon Barocas, Kate Crawford, Aaron Shapiro, and Hanna Wallach. The problem with bias: Allocative versus representational harms in machine learning. In *9th Annual Conference of the Special Interest Group for Computing, Information and Society*, volume 1, 2017. URL <https://meetings.sigcis.org/uploads/6/3/6/8/6368912/program.pdf>.
- Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. The MIT Press, 2023. ISBN 9780262048613. URL <https://mitpress.mit.edu/9780262048613/fairness-and-machine-learning/>.
- Martijn Bartelds, Ananjan Nandi, Moussa Koulako Bala Doumbouya, Dan Jurafsky, Tatsunori Hashimoto, and Karen Livescu. CTC-DRO: Robust Optimization for Reducing Language Disparities in Speech Recognition, 2025. URL <https://arxiv.org/abs/2502.01777>.
- Oguzhan Baser, Ahmet Ege Tanriverdi, Kaan Kale, Sandeep Chinchali, and Sriram Vishwanath. WavShape: Information-Theoretic Speech Representation Learning for Fair and Privacy-Aware Audio Processing. In *Interspeech 2025*, pp. 4873–4877. ISCA, 2025. doi: 10.21437/interspeech.2025-2528. URL <https://doi.org/10.21437/interspeech.2025-2528>.
- Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual Information Neural Estimation. In *International Conference on Machine Learning*, volume 80, pp. 531–540. PMLR, 2018. URL <https://proceedings.mlr.press/v80/belghazi18a.html>.
- Samuel Bell, Mariano Coria Meglioli, Megan Richards, Eduardo Sánchez, Christophe Ropers, Skyler Wang, Adina Williams, Levent Sagun, and Marta R. Costa-jussà. On the Role of Speech Data in Reducing Toxicity Detection Bias. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*,

-
- pp. 1454–1468. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.naacl-long.67. URL <https://doi.org/10.18653/v1/2025.naacl-long.67>.
- Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. LEACE: Perfect linear concept erasure in closed form. In *Advances in Neural Information Processing Systems 36*, pp. 66044–66063. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-2884. URL <https://doi.org/10.52202/075280-2884>.
- Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, 1995. ISSN 1369-7412, 1467-9868. doi: 10.1111/j.2517-6161.1995.tb02031.x. URL <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>.
- Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. Gender in Danger? Evaluating Speech Translation Technology on the MuST-SHE Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6923–6933. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.619. URL <https://doi.org/10.18653/v1/2020.acl-main.619>.
- M. Benzeghiba, R. De Mori, O. Deroo, S. Dupont, T. Erbes, D. Jouvet, L. Fissore, P. Laface, A. Mertins, C. Ris, R. Rose, V. Tyagi, and C. Wellekens. Automatic speech recognition and speech variability: A review. *Speech Communication*, 49(10-11):763–786, 2007. ISSN 0167-6393. doi: 10.1016/j.specom.2007.02.006. URL <https://doi.org/10.1016/j.specom.2007.02.006>.
- Adam Eric Berkowitz and Miriam E. Sweeney. Look Who’s Talking: Voice cloning as tension point between identity and data. *Philosophy & Technology*, 38(4):131, 2025. ISSN 2210-5433, 2210-5441. doi: 10.1007/s13347-025-00965-7. URL <https://doi.org/10.1007/s13347-025-00965-7>.
- Susmita Bhattacharjee, Jagabandhu Mishra, H. S. Shekhawat, and S. R. Mahadeva Prasanna. Fairness of Automatic Speech Recognition in Cleft Lip and Palate Speech, 2025. URL <https://doi.org/10.48550/arxiv.2505.03697>.
- Steven Bird. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 3504–3519. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.313. URL <https://doi.org/10.18653/v1/2020.coling-main.313>.
- Abeba Birhane. Algorithmic injustice: a relational ethics approach. *Patterns*, 2(2):100205, 2021. ISSN 2666-3899. doi: 10.1016/j.patter.2021.100205. URL <https://doi.org/10.1016/j.patter.2021.100205>.
- Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.485. URL <https://doi.org/10.18653/v1/2020.acl-main.485>.
- Brandon M. Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K. D’Mello. Bias and Fairness in Multimodal Machine Learning: A Case Study of Automated Video Interviews. In *Proceedings of the 2021 International Conference on Multimodal Interaction*, pp. 268–277. ACM, 2021. doi: 10.1145/3462244.3479897. URL <https://doi.org/10.1145/3462244.3479897>.
- Josephine Beatrice Skovbo Borre, Malene Gorm Wold, Sara Kjær Rasmussen, and Ilhan Aslan. "How to Explore Biases in Speech Emotion AI with Users?" A Speech-Emotion-Acting Study Exploring Age and Language Biases, 2025. URL <https://arxiv.org/abs/2507.12580>.
- Mondher Bouazizi, Chuheng Zheng, Siyuan Yang, and Tomoaki Ohtsuki. Dementia detection from speech: What if language models are not the answer? *Information*, 15(1):2, 2023. ISSN 2078-2489. doi: 10.3390/info15010002. URL <https://doi.org/10.3390/info15010002>.

-
- Robin N. Brewer, Christina Harrington, and Courtney Heldreth. Envisioning Equitable Speech Technologies for Black Older Adults. In *2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 379–388. ACM, 2023. doi: 10.1145/3593013.3594005. URL <https://doi.org/10.1145/3593013.3594005>.
- Jenný Brynjarsdóttir and Anthony O’Hagan. Learning about physical parameters: the importance of model discrepancy. *Inverse Problems*, 30(11):114007, 2014. ISSN 0266-5611, 1361-6420. doi: 10.1088/0266-5611/30/11/114007. URL <https://doi.org/10.1088/0266-5611/30/11/114007>.
- Niko Brümmer and Johan du Preez. Application-independent evaluation of speaker detection. *Computer Speech & Language*, 20(2-3):230–275, 2006. ISSN 0885-2308. doi: 10.1016/j.csl.2005.08.001. URL <https://doi.org/10.1016/j.csl.2005.08.001>. Odyssey 2004: The speaker and Language Recognition Workshop.
- Xiong Cai, Zhiyong Wu, Kuo Zhong, Bin Su, Dongyang Dai, and Helen Meng. Unsupervised Cross-Lingual Speech Emotion Recognition Using Domain Adversarial Neural Network. In *2021 12th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 1–5. IEEE, 2021. doi: 10.1109/iscslp49672.2021.9362058. URL <https://doi.org/10.1109/iscslp49672.2021.9362058>.
- T. Caliński and J Harabasz. A dendrite method for cluster analysis. *Communications in Statistics*, 3(1):1–27, 1974. ISSN 0090-3272. doi: 10.1080/03610927408827101. URL <https://doi.org/10.1080/03610927408827101>.
- Stephanie Russo Carroll, Ibrahim Garba, Oscar L. Figueroa-Rodríguez, Jarita Holbrook, Raymond Lovett, Simeon Materechera, Mark Parsons, Kay Raseroka, Desi Rodríguez-Lonebear, Robyn Rowe, Rodrigo Sara, Jennifer D. Walker, Jane Anderson, and Maui Hudson. The CARE principles for indigenous data governance. *Data Science Journal*, 19:43, 2020. ISSN 1683-1470. doi: 10.5334/dsj-2020-043. URL <https://doi.org/10.5334/dsj-2020-043>.
- Galo Castillo-López, Gael de Chalendar, and Nasredine Semmar. A Survey of Recent Advances on Turn-taking Modeling in Spoken Dialogue Systems. In *Proceedings of the 15th International Workshop on Spoken Dialogue Systems Technology*, pp. 254–271. Association for Computational Linguistics, 2025. ISBN 979-8-89176-248-0. URL <https://aclanthology.org/2025.iwds-1.27/>.
- Shireen Chand, Shwetha Ajay, and Vaishali Kulkarni. From disparity to fairness: Analyzing challenges and solutions in AI systems for disabled individuals. In *2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMi)*, pp. 1–6. IEEE, 2023. doi: 10.1109/cvmi59935.2023.10465050. URL <https://doi.org/10.1109/cvmi59935.2023.10465050>.
- Ankani Chattoraj, Rupam Acharyya, Shouman Das, Md. Iftekhar Tanveer, and Ehsan Hoque. Fairness in Rating Prediction by Awareness of Verbal and Gesture Quality of Public Speeches, 2020. URL <https://doi.org/10.48550/arxiv.2012.06157>.
- Shivam Chauhan and Ajay Pundhir. Cross-Cultural Bias in Mel-Scale Representations: Evidence and Alternatives from Speech and Music. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2101–2105. IEEE, IEEE, 2026. doi: 10.1109/icassp55912.2026.11462385. URL <https://doi.org/10.1109/icassp55912.2026.11462385>.
- Dazhong Chen, Yi-Cheng Lin, Yuchen Huang, Ziwei Gong, Di Jiang, Zeying Xie, Yi R., and Fung. CantoASR: Prosody-Aware ASR-LALM Collaboration for Low-Resource Cantonese, 2025a. URL <https://doi.org/10.48550/arxiv.2511.04139>.
- Kuan-Yu Chen, Yi-Cheng Lin, Po-Chung Hsieh, Huang-Cheng Chou, Chih-Fan Hsu, Jeng-Lin Li, Hung-yi Lee, and Jian-Jiun Ding. The Binding Effect: Analysis of How Multi-Dimensional Cues Form Gender Bias in Instruction TTS. In *Interspeech 2026*, 2026a. URL <https://arxiv.org/abs/2603.20743>. Accepted for publication.
- Szu-Chi Chen, I-Ning Tsai, Yi-Cheng Lin, Sung-Feng Huang, and Hung yi Lee. MoVE: Translating Laughter and Tears via Mixture of Vocalization Experts in Speech-to-Speech Translation, 2026b. URL <https://arxiv.org/abs/2604.17435>.

-
- Yi-Chen Chen, Zhaojun Yang, Ching-Feng Yeh, Mahaveer Jain, and Michael L. Seltzer. Aipnet: Generative Adversarial Pre-Training of Accent-Invariant Networks for End-To-End Speech Recognition. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6979–6983. IEEE, 2020. doi: 10.1109/icassp40776.2020.9053098. URL <https://doi.org/10.1109/icassp40776.2020.9053098>.
- Zigang Chen, Yuening Zhou, Zhen Wang, Fan Liu, Tao Leng, and Haihua Zhu. A bias evaluation solution for multiple sensitive attribute speech recognition. *Computer Speech & Language*, 93:101787, 2025b. ISSN 0885-2308. doi: 10.1016/j.csl.2025.101787. URL <https://doi.org/10.1016/j.csl.2025.101787>.
- Jian Cheng. Automatic assessment of prosody in high-stakes English tests. In *Interspeech 2011*, pp. 1589–1592. ISCA, 2011. doi: 10.21437/interspeech.2011-477. URL <https://doi.org/10.21437/interspeech.2011-477>.
- Woan-Shiuan Chien and Chi-Chun Lee. Achieving fair speech emotion recognition via perceptual fairness. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023. doi: 10.1109/icassp49357.2023.10094984. URL <https://doi.org/10.1109/icassp49357.2023.10094984>.
- Woan-Shiuan Chien and Chi-Chun Lee. An Investigation of Group versus Individual Fairness in Perceptually Fair Speech Emotion Recognition. In *Interspeech 2024*, pp. 3205–3209. ISCA, 2024. doi: 10.21437/interspeech.2024-461. URL <https://doi.org/10.21437/interspeech.2024-461>.
- Woan-Shiuan Chien and Chi-Chun Lee. Only Subsets Matters: The Effect of Dual Fairness Constraints in Speech Emotion Recognition. *IEEE Transactions on Affective Computing*, 17(2):2340–2353, 2026. ISSN 1949-3045, 2371-9850. doi: 10.1109/taffc.2026.3668866. URL <https://doi.org/10.1109/taffc.2026.3668866>.
- Woan-Shiuan Chien, Shreya G. Upadhyay, and Chi-Chun Lee. Balancing Speaker-Rater Fairness for Gender-Neutral Speech Emotion Recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11861–11865. IEEE, 2024. doi: 10.1109/icassp48485.2024.10447167. URL <https://doi.org/10.1109/icassp48485.2024.10447167>.
- Anna Seo Gyeong Choi and Hoon Choi. 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, 2025. ISSN 3065-8365. doi: 10.1609/aies.v8i1.36574. URL <https://doi.org/10.1609/aies.v8i1.36574>.
- Junhyuk Choi, Ro-hoon Oh, Jihwan Seol, and Bugeun Kim. VoiceBBQ: Investigating Effect of Content and Acoustics in Social Bias of Spoken Language Model. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 28713–28724. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.emnlp-main.1461. URL <https://doi.org/10.18653/v1/2025.emnlp-main.1461>.
- Junhyuk Choi, Jihwan Seol, Nayeon Kim, Chanhee Cho, EunBin Cho, and Bugeun Kim. Acoustic-Based Gender Differentiation in Speech-Aware Language Models, 2025b. URL <https://arxiv.org/abs/2509.21125>.
- Hsing-Hang Chou, Woan-Shiuan Chien, Ya-Tse Wu, and Chi-Chun Lee. An inter-speaker fairness-aware speech emotion regression framework. In *Interspeech 2024*, pp. 3190–3194. ISCA, 2024. doi: 10.21437/interspeech.2024-471. URL <https://doi.org/10.21437/interspeech.2024-471>.
- Oubaïda Chouchane, Christoph Busch, Chiara Galdi, Nicholas Evans, and Massimiliano Todisco. A Comparison of Differential Performance Metrics for the Evaluation of Automatic Speaker Verification Fairness. In *The Speaker and Language Recognition Workshop (Odyssey 2024)*, pp. 209–216. ISCA, 2024. doi: 10.21437/odyssey.2024-30. URL <https://doi.org/10.21437/odyssey.2024-30>.
- Kimberly Christen. Tribal archives, traditional knowledge, and local contexts: Why the “s” matters. *Journal of Western Archives*, 6(1):Article 3, 2015. ISSN 2154-7149. URL <https://digitalcommons.usu.edu/westernarchives/vol6/iss1/3/>.

-
- Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, 2013. ISBN 9781134742707. doi: 10.4324/9780203771587. URL <https://doi.org/10.4324/9780203771587>.
- Pierre Colombo, Nathan Noiry, Guillaume Staerman, and Pablo Piantanida. A Novel Information Theoretic Objective to Disentangle Representations for Fair Classification. In *Findings of the Association for Computational Linguistics: IJCNLP-AACL 2023 (Findings)*, pp. 184–198. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-ijcnlp.17. URL <https://doi.org/10.18653/v1/2023.findings-ijcnlp.17>.
- A. Feder Cooper, Emanuel Moss, Benjamin Laufer, and Helen Nissenbaum. Accountability in an algorithmic society: Relationality, responsibility, and robustness in machine learning. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, FAccT ’22*, pp. 864–876, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393522. doi: 10.1145/3531146.3533150. URL <https://doi.org/10.1145/3531146.3533150>.
- Marta R. Costa-jussà, Christine Basta, and Gerard I. Gállego. Evaluating Gender Bias in Speech Translation. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 2141–2147. European Language Resources Association (ELRA), 2022. doi: 10.63317/586hoxzn9ig8. URL <https://doi.org/10.63317/586hoxzn9ig8>.
- Nick Couldry and Ulises A. Mejias. Data colonialism: Rethinking big data’s relation to the contemporary subject. *Television & New Media*, 20(4):336–349, 2019. ISSN 1527-4764, 1552-8316. doi: 10.1177/1527476418796632. URL <https://doi.org/10.1177/1527476418796632>.
- Wenqian Cui, Dianzhi Yu, Xiaoqi Jiao, Ziqiao Meng, Guangyan Zhang, Qichao Wang, Steven Y. Guo, and Irwin King. Recent Advances in Speech Language Models: A Survey. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13943–13970. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.682. URL <https://doi.org/10.18653/v1/2025.acl-long.682>.
- M.L. Cummings. Automation bias in intelligent time critical decision support systems. In *Decision Making in Aviation*, pp. 289–294. Routledge, 2017. ISBN 9781315095080. doi: 10.4324/9781315095080-17. URL <https://doi.org/10.4324/9781315095080-17>.
- Jay L. Cunningham, Adinawa Adjagbodjou, Jeffrey Basoah, Jainaba Jawara, Kowe Kadoma, and Aaleyah Lewis. Toward Responsible ASR for African American English Speakers: A Scoping Review of Bias and Equity in Speech Technology. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 8(1):665–678, 2025. ISSN 3065-8365. doi: 10.1609/aies.v8i1.36580. URL <https://doi.org/10.1609/aies.v8i1.36580>.
- Zhengjia Dan, Yue Zhao, Xiaojun Bi, Licheng Wu, and Qiang Ji. Multi-Task Transformer with Adaptive Cross-Entropy Loss for Multi-Dialect Speech Recognition. *Entropy*, 24(10):1429, 2022. ISSN 1099-4300. doi: 10.3390/e24101429. URL <https://doi.org/10.3390/e24101429>.
- Mayukh Das and Wolf-Tilo Balke. Quantifying Bias from Decoding Techniques in Natural Language Generation. In *COLING*, pp. 1311–1323, 2022. URL <https://aclanthology.org/2022.coling-1.112>.
- Nilaksh Das, Sravan Bodapati, Monica Sunkara, Sundararajan Srinivasan, and Duen Horng Chau. Best of both worlds: Robust accented speech recognition with adversarial transfer learning. In *Interspeech 2021*, pp. 1314–1318. ISCA, 2021. doi: 10.21437/interspeech.2021-1888. URL <https://doi.org/10.21437/interspeech.2021-1888>.
- David L. Davies and Donald W. Bouldin. A Cluster Separation Measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-1(2):224–227, 1979. ISSN 0162-8828, 2160-9292. doi: 10.1109/tpami.1979.4766909. URL <https://doi.org/10.1109/tpami.1979.4766909>.
- Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. Bias in Bios: A Case Study

-
- of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 120–128. ACM, 2019. doi: 10.1145/3287560.3287572. URL <https://doi.org/10.1145/3287560.3287572>.
- Tiago de Freitas Pereira and Sebastien Marcel. Fairness in Biometrics: A Figure of Merit to Assess Biometric Verification Systems. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4(1):19–29, 2022. ISSN 2637-6407. doi: 10.1109/tbiom.2021.3102862. URL <https://doi.org/10.1109/tbiom.2021.3102862>.
- Sepehr Dehdashtian, Ruozhen He, Yi Li, Guha Balakrishnan, Nuno Vasconcelos, Vicente Ordonez, and Vishnu Naresh Boddeti. Fairness and Bias Mitigation in Computer Vision: A Survey, 2024. URL <https://doi.org/10.48550/arxiv.2408.02464>.
- Miguel Del Río, Corey Miller, Ján Profant, Jennifer Drexler-Fox, Quinn Mcnamara, Nishchal Bhandari, Natalie Delworth, Ilya Pirkin, Migüel Jetté, Shipra Chandra, Peter Ha, and Ryan Westerman. Accents in speech recognition through the lens of a world englishes evaluation set. *Research in Language*, 21(3): 225–244, 2023. ISSN 1731-7533. doi: 10.18778/1731-7533.21.3.02. URL <https://doi.org/10.18778/1731-7533.21.3.02>.
- Chengxi Deng, Xurong Xie, Shujie Hu, Mengzhe Geng, Yicong Jiang, Jiankun Zhao, Jiajun Deng, Guinan Li, Youjun Chen, Huimeng Wang, Haoning Xu, Mingyu Cui, and Xunying Liu. MOPSA: Mixture of Prompt-Experts Based Speaker Adaptation for Elderly Speech Recognition. In *Interspeech 2025*, pp. 4933–4937. ISCA, 2025. doi: 10.21437/interspeech.2025-1263. URL <https://doi.org/10.21437/interspeech.2025-1263>.
- PRANAV DHERAM, Murugesan Ramakrishnan, Anirudh Raju, I-Fan Chen, Brian King, Katherine Powell, Melissa Saboowala, Karan Shetty, and Andreas Stolcke. Toward Fairness in Speech Recognition: Discovery and Mitigation of Performance Disparities. In *Interspeech 2022*, pp. 1268–1272. ISCA, 2022. doi: 10.21437/interspeech.2022-10816. URL <https://doi.org/10.21437/interspeech.2022-10816>.
- Roel Dobbe, Sarah Dean, Thomas Gilbert, and Nitin Kohli. A broader view on bias in automated decision-making: Reflecting on epistemology and dynamics, 2018. URL <https://doi.org/10.48550/arxiv.1807.00553>.
- Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226. ACM, 2012. doi: 10.1145/2090236.2090255. URL <https://doi.org/10.1145/2090236.2090255>.
- Nina Sun Eidsheim. *The race of sound: Listening, timbre, and vocality in African American music*. Duke University Press, 2019. ISBN 9781478090359. doi: 10.1215/9781478090359. URL <https://doi.org/10.1215/9781478090359>.
- Yanai Elazar and Yoav Goldberg. Adversarial Removal of Demographic Attributes from Text Data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 11–21. Association for Computational Linguistics, 2018. doi: 10.18653/v1/d18-1002. URL <https://doi.org/10.18653/v1/d18-1002>.
- Hend ElGhazaly, Bahman Mirheidari, Heidi Christensen, and Nafise Sadat Moosavi. Fairness in Automatic Speech Recognition Isn’t a One-Size-Fits-All. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19169–19178. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.1044. URL <https://doi.org/10.18653/v1/2025.findings-emnlp.1044>.
- Alaa Elobaid, Nathan Ramoly, Lara Younes, Symeon Papadopoulos, Eirini Ntoutsis, and Ioannis Kompatsiaris. Sum of Group Error Differences: A Critical Examination of Bias Evaluation in Biometric Verification and a Dual-Metric Measure. In *2024 IEEE 18th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–9. IEEE, 2024. doi: 10.1109/fg59268.2024.10582012. URL <https://doi.org/10.1109/fg59268.2024.10582012>.

-
- Chris Emezue, The NaijaVoices Community, Busayo Awobade, Abraham Toluwase Owodunni, Handel Emezue, Gloria Monica Tobechukwu Emezue, Nefertiti Nneoma Emezue, Sewade Ogun, Bunmi Akinremi, David Ifeoluwa Adelani, and Chris Pal. The NaijaVoices Dataset: Cultivating Large-Scale, High-Quality, Culturally-Rich Speech Data for African Languages. In *Interspeech 2025*, pp. 1338–1342. ISCA, 2025. doi: 10.21437/interspeech.2025-1104. URL <https://doi.org/10.21437/interspeech.2025-1104>.
- Mariel Estevez and Luciana Ferrer. Study on the Fairness of Speaker Verification Systems Across Accent and Gender Groups. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023. doi: 10.1109/icassp49357.2023.10095150. URL <https://doi.org/10.1109/icassp49357.2023.10095150>.
- Mariel Estevez, Cyntia Bonomi, Dayana Ribas, Alfonso Ortega, and Luciana Ferrer. Beyond Global Metrics: A Fairness Analysis for Interpretable Voice Disorder Detection Systems. *Journal of Voice*, 2025. ISSN 0892-1997. doi: 10.1016/j.jvoice.2025.11.037. URL <https://doi.org/10.1016/j.jvoice.2025.11.037>.
- Kelechi Ezema, Chelsea Chandler, Rosy Southwell, Niranjana Cholendiran, and Sidney D’Mello. “It feels like we’re not meeting the criteria”: Examining and Mitigating the Cascading Effects of Bias in Automatic Speech Recognition in Spoken Language Interfaces. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. ACM, 2025. doi: 10.1145/3706598.3714059. URL <https://doi.org/10.1145/3706598.3714059>.
- Simone Facchiano, Giorgio Strano, Donato Crisostomi, Irene Tallini, Tommaso Mencattini, Fabio Galasso, and Emanuele Rodolà. Activation Patching for Interpretable Steering in Music Generation, 2025. URL <https://doi.org/10.48550/arxiv.2504.04479>.
- Yasaman Fatapour, Jamil S. Samaan, Inclusive AI Research Group, and Nicholas P Tatonetti. Accents Still Confuse AI: Systematic Errors in Speech Transcription and LLM-Based Remedies, 2025. URL <https://doi.org/10.1101/2025.08.29.25333548>.
- Siyuan Feng, Olya Kudina, Bence Mark Halpern, and Odette Scharenborg. Quantifying Bias in Automatic Speech Recognition, 2021. URL <https://arxiv.org/abs/2103.15122>.
- Siyuan Feng, Bence Mark Halpern, Olya Kudina, and Odette Scharenborg. Towards inclusive automatic speech recognition. *Computer Speech & Language*, 84:101567, 2024. ISSN 0885-2308. doi: 10.1016/j.csl.2023.101567. URL <https://doi.org/10.1016/j.csl.2023.101567>.
- Tiantian Feng and Shrikanth Narayanan. PEFT-SER: On the Use of Parameter Efficient Transfer Learning Approaches For Speech Emotion Recognition Using Pre-trained Speech Models. In *2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII)*, pp. 1–8. IEEE, 2023. doi: 10.1109/acii59096.2023.10388152. URL <https://doi.org/10.1109/acii59096.2023.10388152>.
- Tiantian Feng, Rajat Hebbar, Nicholas Mehlman, Xuan Shi, Aditya Kommineni, and Shrikanth Narayanan. A review of speech-centric trustworthy machine learning: Privacy, safety, and fairness. *APSIPA Transactions on Signal and Information Processing*, 12(3), 2023. ISSN 2048-7703. doi: 10.1561/116.00000084. URL <https://doi.org/10.1561/116.00000084>.
- Gianni Fenu and Mirko Marras. Demographic Fairness in Multimodal Biometrics: A Comparative Analysis on Audio-Visual Speaker Recognition Systems. *Procedia Computer Science*, 198:249–254, 2022. ISSN 1877-0509. doi: 10.1016/j.procs.2021.12.236. URL <https://doi.org/10.1016/j.procs.2021.12.236>. 12th International Conference on Emerging Ubiquitous Systems and Pervasive Networks / 11th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare.
- Gianni Fenu, Hicham Lafhouli, and Mirko Marras. Exploring Algorithmic Fairness in Deep Speaker Verification. In *Lecture Notes in Computer Science*, pp. 77–93. Springer International Publishing, 2020a. ISBN 9783030588106, 9783030588113. doi: 10.1007/978-3-030-58811-3_6. URL https://doi.org/10.1007/978-3-030-58811-3_6.

-
- Gianni Fenu, Giacomo Medda, Mirko Marras, and Giacomo Meloni. Improving Fairness in Speaker Recognition. In *Proceedings of the 2020 European Symposium on Software Engineering*, pp. 129–136. ACM, 2020b. doi: 10.1145/3393822.3432325. URL <https://doi.org/10.1145/3393822.3432325>.
- Gianni Fenu, Mirko Marras, Giacomo Medda, and Giacomo Meloni. Fair Voice Biometrics: Impact of Demographic Imbalance on Group Fairness in Speaker Recognition. In *Interspeech 2021*, pp. 1892–1896. ISCA, 2021. doi: 10.21437/interspeech.2021-1857. URL <https://doi.org/10.21437/interspeech.2021-1857>.
- Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems*, 14(3):330–347, 1996. ISSN 1046-8188, 1558-2868. doi: 10.1145/230538.230561. URL <https://doi.org/10.1145/230538.230561>.
- Dennis Fucci, Marco Gaido, Matteo Negri, Mauro Cettolo, and Luisa Bentivogli. No Pitch Left Behind: Addressing Gender Unbalance In Automatic Speech Recognition Through Pitch Manipulation. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023. doi: 10.1109/asru57964.2023.10389767. URL <https://doi.org/10.1109/asru57964.2023.10389767>.
- Marco Gaido, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. How To Build Competitive Multi-gender Speech Translation Models For Controlling Speaker Gender Translation, 2023. URL <https://doi.org/10.48550/arxiv.2310.15114>.
- Elroy Galbraith, Chadwick Sutherland, and Donahue Morgan. TRIDENT: A Redundant Architecture for Caribbean-Accented Emergency Speech Triage, 2025. URL <https://doi.org/10.48550/arxiv.2512.10741>.
- Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3):1097–1179, 2024. ISSN 0891-2017, 1530-9312. doi: 10.1162/coli_a_00524. URL https://doi.org/10.1162/coli_a_00524.
- Yaroslav Ganin and Victor Lempitsky. Unsupervised Domain Adaptation by Backpropagation, 2014. URL <https://doi.org/10.48550/arxiv.1409.7495>.
- Heting Gao, Xiaoxuan Wang, Sunghun Kang, Rusty Mina, Dias Issa, John Harvill, Leda Sari, Mark Hasegawa-Johnson, and Chang D. Yoo. Seamless equal accuracy ratio for inclusive CTC speech recognition. *Speech Communication*, 136:76–83, 2022. ISSN 0167-6393. doi: 10.1016/j.specom.2021.11.004. URL <https://doi.org/10.1016/j.specom.2021.11.004>.
- Yuan Gao, JiaXing Liu, Longbiao Wang, and Jianwu Dang. Domain-Adversarial Autoencoder with Attention Based Feature Level Fusion for Speech Emotion Recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6314–6318. IEEE, 2021. doi: 10.1109/icassp39728.2021.9413710. URL <https://doi.org/10.1109/icassp39728.2021.9413710>.
- Shefali Garg, Zhouyuan Huo, Khe Chai Sim, Suzan Schwartz, Mason Chua, Alëna Aksënova, Tsendsuren Munkhdalai, Levi King, Darryl Wright, Zion Mengesha, Dongseong Hwang, Tara Sainath, Françoise Beaufays, and Pedro Moreno Mengibar. Improving Speech Recognition for African American English with Audio Classification. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12356–12360. IEEE, 2024. doi: 10.1109/icassp48485.2024.10447116. URL <https://doi.org/10.1109/icassp48485.2024.10447116>.
- Mahault Garnerin, Solange Rossato, and Laurent Besacier. Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance. In *Proceedings of the 1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, pp. 3–9. ACM, 2019. doi: 10.1145/3347449.3357480. URL <https://doi.org/10.1145/3347449.3357480>.
- Stuart Geman, Elie Bienenstock, and René Doursat. Neural Networks and the Bias/Variance Dilemma. *Neural Computation*, 4(1):1–58, 1992. ISSN 0899-7667, 1530-888X. doi: 10.1162/neco.1992.4.1.1. URL <https://doi.org/10.1162/neco.1992.4.1.1>.

-
- Mengzhe Geng, Xurong Xie, Rongfeng Su, Jianwei Yu, Zengrui Jin, Tianzi Wang, Shujie Hu, Zi Ye, Helen Meng, and Xunying Liu. On-the-Fly Feature Based Rapid Speaker Adaptation for Dysarthric and Elderly Speech Recognition. In *INTERSPEECH 2023*, pp. 1753–1757. ISCA, 2023. doi: 10.21437/interspeech.2023-301. URL <https://doi.org/10.21437/interspeech.2023-301>.
- Mengzhe Geng, Xurong Xie, Jiajun Deng, Zengrui Jin, Guinan Li, Tianzi Wang, Shujie Hu, Zhaoqing Li, Helen Meng, and Xunying Liu. Homogeneous Speaker Features for on-the-Fly Dysarthric and Elderly Speaker Adaptation and Speech Recognition. *IEEE Transactions on Audio, Speech and Language Processing*, 33: 1689–1705, 2025. ISSN 2998-4173. doi: 10.1109/taslpro.2025.3547217. URL <https://doi.org/10.1109/taslpro.2025.3547217>.
- Corrado Gini. *Variabilità e mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche.[Fasc. I.]*. Tipografia di Paolo Cuppini, 1912. URL <https://search.worldcat.org/title/83150327>.
- Jose Giraldo, Alex Peiró-Lilja, Carme Armentano-Oller, Rodolfo Zevallos, and Cristina España-Bonet. Evaluating Speech Enhancement Performance Across Demographics and Language. In *Interspeech 2025*, pp. 1353–1357. ISCA, 2025. doi: 10.21437/interspeech.2025-1760. URL <https://doi.org/10.21437/interspeech.2025-1760>.
- Usman Gohar and Lu Cheng. A survey on intersectional fairness in machine learning: Notions, mitigation, and challenges. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 6619–6627. International Joint Conferences on Artificial Intelligence Organization, 2023. doi: 10.24963/ijcai.2023/742. URL <https://doi.org/10.24963/ijcai.2023/742>.
- Youdi Gong, Guangzhen Liu, Yunzhi Xue, Rui Li, and Lingzhong Meng. A survey on dataset quality in machine learning. *Information and Software Technology*, 162:107268, 2023. ISSN 0950-5849. doi: 10.1016/j.infsof.2023.107268. URL <https://doi.org/10.1016/j.infsof.2023.107268>.
- Cristina Gorrostieta, Reza Lotfian, Kye Taylor, Richard Brutti, and John Kane. Gender De-Biasing in Speech Emotion Recognition. In *Interspeech 2019*, pp. 2823–2827. ISCA, 2019. doi: 10.21437/interspeech.2019-1708. URL <https://doi.org/10.21437/interspeech.2019-1708>.
- Patrick Grother. Face recognition vendor test (FRVT) part 8: Summarizing demographic differentials. Technical report, National Institute of Standards and Technology, 2022. URL <https://doi.org/10.6028/NIST.IR.8429.ipd>.
- Xiangming Gu, Wei Zeng, and Ye Wang. Elucidate Gender Fairness in Singing Voice Transcription. In *Proceedings of the 31st ACM International Conference on Multimedia*, pp. 8760–8769. ACM, 2023. doi: 10.1145/3581783.3612272. URL <https://doi.org/10.1145/3581783.3612272>.
- Kashaf Gulzar, Korbinian Riedhammer, Elmar Nöth, Andreas K. Maier, and Paula Andrea Pérez-Toro. Bias and Fairness in Self-Supervised Acoustic Representations for Cognitive Impairment Detection, 2026. URL <https://arxiv.org/abs/2603.02937>.
- Yasaman Haghbin, Sina Rashidi, Ali Zolnour, and Maryam Zolnoori. The Voice of Equity: A Systematic Evaluation of Bias Mitigation Techniques for Speech-Based Cognitive Impairment Detection Across Architectures and Demographics, 2026. URL <https://arxiv.org/abs/2601.16989>.
- Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. Gender recognition or gender reductionism? The social implications of embedded gender recognition systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–13. ACM, 2018. doi: 10.1145/3173574.3173582. URL <https://doi.org/10.1145/3173574.3173582>.
- Tao Han, Hantao Huang, Ziang Yang, and Wei Han. Supervised contrastive learning for accented speech recognition, 2021. URL <https://doi.org/10.48550/arxiv.2107.00921>.
- Alex Hanna, Remi Denton, Andrew Smart, and Jamila Smith-Loud. Towards a critical race methodology in algorithmic fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 501–512. ACM, 2020. doi: 10.1145/3351095.3372826. URL <https://doi.org/10.1145/3351095.3372826>.

-
- Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29*, pp. 3315–3323, 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/6a9659feb1216f14f7384ba499518b38-Abstract.html.
- Camille Harris, Chijioke Mgbahurike, Neha Kumar, and Diyi Yang. Modeling Gender and Dialect Bias in Automatic Speech Recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 15166–15184. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.890. URL <https://doi.org/10.18653/v1/2024.findings-emnlp.890>.
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer New York, 2009. ISBN 9780387848570, 9780387848587. doi: 10.1007/978-0-387-84858-7. URL <https://doi.org/10.1007/978-0-387-84858-7>.
- Liu He, Yuanchao Li, Rui Feng, XinRan Han, Yin-Long Liu, Yuwei Yang, Zude Zhu, and Jiahong Yuan. Exploring Gender Bias in Alzheimer’s Disease Detection: Insights from Mandarin and Greek Speech Perception. In *Communications in Computer and Information Science*, pp. 213–224. Springer Nature Singapore, 2026. ISBN 9789819553815, 9789819553822. doi: 10.1007/978-981-95-5382-2_17. URL https://doi.org/10.1007/978-981-95-5382-2_17.
- Zexue He, Bodhisattwa Prasad Majumder, and Julian McAuley. Detect and Perturb: Neutral Rewriting of Biased and Sensitive Text via Gradient-based Decoding. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pp. 4173–4181. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.findings-emnlp.352. URL <https://doi.org/10.18653/v1/2021.findings-emnlp.352>.
- Louis Hickman, Markus Langer, Rachel M. Saef, and Louis Tay. Automated speech recognition bias in personnel selection: The case of automatically scored job interviews. *Journal of Applied Psychology*, 110(6):846–858, 2025. ISSN 1939-1854, 0021-9010. doi: 10.1037/apl0001247. URL <https://doi.org/10.1037/apl0001247>.
- Arthur Hinsvark, Natalie Delworth, Miguel Del Rio, Quinten McNamara, Joshua Dong, Ryan Westerman, Michelle Huang, Joseph Palakapilly, Jennifer Drexler, Ilya Pirkin, Nishchal Bhandari, and Miguel Jette. Accented speech recognition: A survey, 2021. URL <https://doi.org/10.48550/arxiv.2104.10747>.
- Tobias Hodgson and Enrico Coiera. Risks and benefits of speech recognition for clinical documentation: a systematic review. *Journal of the American Medical Informatics Association*, 23(e1):e169–e179, 2016. ISSN 1527-974X, 1067-5027. doi: 10.1093/jamia/ocv152. URL <https://doi.org/10.1093/jamia/ocv152>.
- Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. What Do Compressed Deep Neural Networks Forget?, 2019. URL <https://arxiv.org/abs/1911.05248>.
- John J. Howard, Eli J. Laird, Rebecca E. Rubin, Yevgeniy B. Sirotnin, Jerry L. Tipton, and Arun R. Vemury. Evaluating Proposed Fairness Models for Face Recognition Algorithms. In *Lecture Notes in Computer Science*, pp. 431–447. Springer Nature Switzerland, 2023. ISBN 9783031376597, 9783031376603. doi: 10.1007/978-3-031-37660-3_31. URL https://doi.org/10.1007/978-3-031-37660-3_31.
- Wei-Ning Hsu, Yu Zhang, and James Glass. Unsupervised learning of disentangled and interpretable representations from sequential data. *Neural Information Processing Systems*, 30:1878–1889, 2017. URL <https://papers.nips.cc/paper/2017/file/0a0a0c8aaa00ade50f74a3f0ca981ed7-Paper.pdf>.
- Wei-Tung Hsu, Chin-Po Chen, and Chi-Chun Lee. Concealing Medical Condition by Node Toggling in ASR for Dementia Patients. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12496–12500. IEEE, 2024. doi: 10.1109/icassp48485.2024.10446570. URL <https://doi.org/10.1109/icassp48485.2024.10446570>.
- Shujie Hu, Xurong Xie, Mengzhe Geng, Jiajun Deng, Zengrui Jin, Tianzi Wang, Mingyu Cui, Guinan Li, Zhaoqing Li, Helen Meng, and Xunying Liu. Structured Speaker-Deficiency Adaptation of Foundation Models for Dysarthric and Elderly Speech Recognition, 2024a. URL <https://arxiv.org/abs/2412.18832>.

-
- Shujie Hu, Xurong Xie, Mengzhe Geng, Zengrui Jin, Jiajun Deng, Guinan Li, Yi Wang, Mingyu Cui, Tianzi Wang, Helen Meng, and Xunying Liu. Self-Supervised ASR Models and Features for Dysarthric and Elderly Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3561–3575, 2024b. ISSN 2329-9290, 2329-9304. doi: 10.1109/taslp.2024.3422839. URL <https://doi.org/10.1109/taslp.2024.3422839>.
- Shujie Hu, Xurong Xie, Mengzhe Geng, Jiajun Deng, Huimeng Wang, Guinan Li, Chengxi Deng, Tianzi Wang, Mingyu Cui, Helen Meng, and Xunying Liu. On-the-fly Routing for Zero-shot MoE Speaker Adaptation of Speech Foundation Models for Dysarthric Speech Recognition. In *Interspeech 2025*, pp. 4608–4612. ISCA, 2025a. doi: 10.21437/interspeech.2025-721. URL <https://doi.org/10.21437/interspeech.2025-721>.
- Xinhui Hu, Smit Desai, Morgan Lundy, and Jessie Chin. Beyond Functionality: Co-Designing Voice User Interfaces for Older Adults’ Well-being. In *Proceedings of the 7th ACM Conference on Conversational User Interfaces*, pp. 1–15. ACM, 2025b. doi: 10.1145/3719160.3736613. URL <https://doi.org/10.1145/3719160.3736613>.
- Elwin Huaman, Wendi Huaman, Jorge Luis Huaman, and Ninfa Quispe. Quechua Speech Datasets in Common Voice: The Case of Puno Quechua. In *Communications in Computer and Information Science*, pp. 184–193. Springer Nature Switzerland, 2026. ISBN 9783032203212, 9783032203229. doi: 10.1007/978-3-032-20322-9_13. URL https://doi.org/10.1007/978-3-032-20322-9_13.
- Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T. Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, William Chen, Chih-Kai Yang, Wenze Ren, Xuanjun Chen, Chi-Yuan Hsiao, Puyuan Peng, Shih-Heng Wang, Chun-Yi Kuan, Ke-Han Lu, Kai-Wei Chang, Fabian Ritter-Gutierrez, Kuan-Po Huang, Siddhant Arora, You-Kuan Lin, Ming To Chuang, Eunjung Yeo, Calvin Chang, Chung-Ming Chien, Kwanghee Choi, Jun-You Wang, Cheng-Hsiu Hsieh, Yi-Cheng Lin, Chee-En Yu, I-Hsiang Chiu, Heitor R. Guimarães, Jionghao Han, Tzu-Quan Lin, Tzu-Yuan Lin, Homu Chang, Ting-Wu Chang, Chun Wei Chen, Shou-Jen Chen, Yu-Hua Chen, Hsi-Chun Cheng, Kunal Dhawan, Jia-Lin Fang, Shi-Xin Fang, Kuan-Yu Fang Chiang, Chi An Fu, Hsien-Fu Hsiao, Ching Yu Hsu, Shao-Syuan Huang, Lee Chen Wei, Hsi-Che Lin, Hsuan-Hao Lin, Hsuan-Ting Lin, Jian-Ren Lin, Ting-Chun Liu, Li-Chun Lu, Tsung-Min Pai, Ankita Pasad, Shih-Yun Shan Kuan, Suwon Shon, Yuxun Tang, Yun-Shao Tsai, Jui-Chiang Wei, Tzu-Chieh Wei, Chengxi Wu, Dien-Ruei Wu, Chao-Han Huck Yang, Chieh-Chi Yang, Jia Qi Yip, Shao-Xiang Yuan, Vahid Noroozi, Zhehuai Chen, Haibin Wu, Karen Livescu, David Harwath, Shinji Watanabe, and Hung-yi Lee. Dynamic-SUPERB Phase-2: A Collaboratively Expanding Benchmark for Measuring the Capabilities of Spoken Language Models with 180 Tasks, 2024. URL <https://doi.org/10.48550/arxiv.2411.05361>.
- Sung-Feng Huang, Chyi-Jiunn Lin, Da-Rong Liu, Yi-Chen Chen, and Hung-yi Lee. Meta-TTS: Meta-learning for few-shot speaker adaptive text-to-speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30:1558–1571, 2022a. ISSN 2329-9290, 2329-9304. doi: 10.1109/taslp.2022.3167258. URL <https://doi.org/10.1109/taslp.2022.3167258>.
- Wen-Chin Huang. Advancing Speech Quality Assessment Through Scientific Challenges and Open-source Activities. In *2025 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pp. 2552–2557. IEEE, 2025. doi: 10.1109/apsipaasc65261.2025.11249197. URL <https://doi.org/10.1109/apsipaasc65261.2025.11249197>.
- Yu-Lin Huang, Bo-Hao Su, Y.-W. Peter Hong, and Chi-Chun Lee. An Attribute-Aligned Strategy for Learning Speech Representation. In *Interspeech 2021*, pp. 1179–1183. ISCA, 2021. doi: 10.21437/interspeech.2021-1341. URL <https://doi.org/10.21437/interspeech.2021-1341>.
- Yu-Lin Huang, Bo-Hao Su, Y.-W. Peter Hong, and Chi-Chun Lee. An Attention-Based Method for Guiding Attribute-Aligned Speech Representation Learning. In *Interspeech 2022*, pp. 5030–5034. ISCA, 2022b. doi: 10.21437/interspeech.2022-10419. URL <https://doi.org/10.21437/interspeech.2022-10419>.
- J. J. Humphries, P. C. Woodland, and D. Pearce. Using accent-specific pronunciation modelling for robust speech recognition. In *4th International Conference on Spoken Language Processing (ICSLP 1996)*, pp. 2324–2327. ISCA, 1996. doi: 10.21437/icslp.1996-588. URL <https://doi.org/10.21437/icslp.1996-588>.

-
- Wiebke Hutiri, Orestis Papakyriakopoulos, and Alice Xiang. Not My Voice! A Taxonomy of Ethical and Safety Harms of Speech Generators. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 359–376. ACM, 2024a. doi: 10.1145/3630106.3658911. URL <https://doi.org/10.1145/3630106.3658911>.
- Wiebke Hutiri, Tanvina Patel, Aaron Yi Ding, and Odette Scharenborg. As Biased as You Measure: Methodological Pitfalls of Bias Evaluations in Speaker Verification Research. In *Interspeech 2024*, pp. 4268–4272. ISCA, 2024b. doi: 10.21437/interspeech.2024-1158. URL <https://doi.org/10.21437/interspeech.2024-1158>.
- Wiebke Toussaint Hutiri and Aaron Yi Ding. Bias in Automated Speaker Recognition. In *2022 ACM Conference on Fairness Accountability and Transparency*, pp. 230–247. ACM, 2022. doi: 10.1145/3531146.3533089. URL <https://doi.org/10.1145/3531146.3533089>.
- Ryo Imaizumi, Ryo Masumura, Sayaka Shiota, and Hitoshi Kiya. End-to-end japanese multi-dialect speech recognition and dialect identification with multi-task learning. *APSIPA Transactions on Signal and Information Processing*, 11(1):1–23, 2022. ISSN 2048-7703, 2048-7703. doi: 10.1561/116.00000045. URL <https://doi.org/10.1561/116.00000045>.
- Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 375–385. ACM, 2021. doi: 10.1145/3442188.3445901. URL <https://doi.org/10.1145/3442188.3445901>.
- Maliha Jahan, Laureano Moro-Velazquez, Thomas Thebaud, Najim Dehak, and Jesús Villalba. Model-Based Fairness Metric for Speaker Verification. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–7. IEEE, 2023. doi: 10.1109/asru57964.2023.10389804. URL <https://doi.org/10.1109/asru57964.2023.10389804>.
- Navdeep Jain and Hongcheng Wang. Inclusive Speaker Verification with Adaptive thresholding, 2021. URL <https://arxiv.org/abs/2111.05501>.
- Iris Dania Jimenez and Christoph Kern. Dialect and Gender Bias in YouTube’s Spanish Captioning System, 2026. URL <https://arxiv.org/abs/2602.24002>.
- Minho Jin, Chelsea Ju, Zeya Chen, Yi Chieh Liu, Jasha Droppo, and Andreas Stolcke. Adversarial Reweighting for Speaker Verification Fairness. In *Interspeech 2022*, pp. 4800–4804. ISCA, 2022. doi: 10.21437/interspeech.2022-10948. URL <https://doi.org/10.21437/interspeech.2022-10948>.
- Bipasha Kashyap and Pubudu N. Pathirana. Geometric Analysis of Speech Representation Spaces: Topological Disentanglement and Confound Detection, 2026. URL <https://arxiv.org/abs/2602.20823>.
- Chamara Kasun, Chung Soo Ahn, Jagath Rajapakse, Zhiping Lin, and Guang-Bin Huang. Discriminative Adversarial Learning for Speaker Independent Emotion Recognition. In *Interspeech 2022*, pp. 4975–4979. ISCA, 2022. doi: 10.21437/interspeech.2022-285. URL <https://doi.org/10.21437/interspeech.2022-285>.
- Vinaya Sree Katamneni, Aakash Varma Nadimpalli, and Ajita Rattani. Demographic Fairness and Accountability of Audio- and Video-Based Unimodal and Bi-modal Deepfake Detectors. In *Face Recognition Across the Imaging Spectrum*, pp. 205–231. Springer Nature Singapore, 2024. ISBN 9789819720583, 9789819720590. doi: 10.1007/978-981-97-2059-0_8. URL https://doi.org/10.1007/978-981-97-2059-0_8.
- Navdeep Kaur and Parminder Singh. Conventional and contemporary approaches used in text to speech synthesis: A review. *Artificial Intelligence Review*, 56(7):5837–5880, 2023. ISSN 0269-2821, 1573-7462. doi: 10.1007/s10462-022-10315-0. URL <https://doi.org/10.1007/s10462-022-10315-0>.
- Christina Kauschke, Daniela Bahn, Michael Vesker, and Gudrun Schwarzer. The role of emotional valence for the processing of facial and verbal stimuli—positivity or negativity bias? *Frontiers in Psychology*, 10:1654, 2019. ISSN 1664-1078. doi: 10.3389/fpsyg.2019.01654. URL <https://doi.org/10.3389/fpsyg.2019.01654>.

-
- Finnian Kelly and Naomi Harte. Effects of Long-Term Ageing on Speaker Verification. In *Lecture Notes in Computer Science*, pp. 113–124. Springer Berlin Heidelberg, 2011. ISBN 9783642195297, 9783642195303. doi: 10.1007/978-3-642-19530-3_11. URL https://doi.org/10.1007/978-3-642-19530-3_11.
- Finnian Kelly, Rahim Saeidi, Naomi Harte, and David A. van Leeuwen. Effect of long-term ageing on i-vector speaker verification. In *Interspeech 2014*. ISCA, 2014. doi: 10.21437/interspeech.2014-18. URL <https://doi.org/10.21437/interspeech.2014-18>.
- Os Keyes. The Misgendering Machines: Trans/HCI Implications of Automatic Gender Recognition. *Proceedings of the ACM on Human-Computer Interaction*, 2(CSCW):1–22, 2018. ISSN 2573-0142. doi: 10.1145/3274357. URL <https://doi.org/10.1145/3274357>.
- Changhun Kim, Joonhyung Park, Hajin Shim, and Eunho Yang. SGEM: Test-Time Adaptation for Automatic Speech Recognition via Sequential-Level Generalized Entropy Minimization. In *INTERSPEECH 2023*, pp. 3367–3371. ISCA, 2023a. doi: 10.21437/interspeech.2023-1282. URL <https://doi.org/10.21437/interspeech.2023-1282>.
- Changwoo Kim, Jinho Choi, Jongyeon Yoon, Daehun Yoo, and Woojin Lee. Fairness-Aware Multimodal Learning in Automatic Video Interview Assessment. *IEEE Access*, 11:122677–122693, 2023b. ISSN 2169-3536. doi: 10.1109/access.2023.3325891. URL <https://doi.org/10.1109/access.2023.3325891>.
- Jongsuk Kim, Jaemyung Yu, Minchan Kwon, and Junmo Kim. FairASR: Fair Audio Contrastive Learning for Automatic Speech Recognition. In *Interspeech 2025*, pp. 3878–3882. ISCA, 2025a. doi: 10.21437/interspeech.2025-2590. URL <https://doi.org/10.21437/interspeech.2025-2590>.
- June-Woo Kim, Haram Yoon, Wonkyo Oh, Dawoon Jung, Sung-Hoon Yoon, Dae-Jin Kim, Dong-Ho Lee, Sang-Yeol Lee, and Chan-Mo Yang. Domain adversarial training for mitigating gender bias in speech-based mental health detection. In *2025 47th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 1–7. IEEE, 2025b. doi: 10.1109/embc58623.2025.11253190. URL <https://doi.org/10.1109/embc58623.2025.11253190>.
- Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the Effects of RLHF on LLM Generalisation and Diversity. In *ICLR*, 2024. URL <https://openreview.net/forum?id=PX3FAVHJT>.
- Joseph F. Klein and Stephen B. Hood. The impact of stuttering on employment opportunities and job performance. *Journal of Fluency Disorders*, 29(4):255–273, 2004. ISSN 0094-730X. doi: 10.1016/j.jfludis.2004.08.001. URL <https://doi.org/10.1016/j.jfludis.2004.08.001>.
- Philipp Klumpp, Pooja Chitkara, Leda Sari, Prashant Serai, Jilong Wu, Irina-Elena Veliche, Rongqing Huang, and Qing He. Synthetic Cross-accent Data Augmentation for Automatic Speech Recognition, 2023. URL <https://arxiv.org/abs/2303.00802>.
- Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Touns, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689, 2020. ISSN 0027-8424, 1091-6490. doi: 10.1073/pnas.1915768117. URL <https://doi.org/10.1073/pnas.1915768117>.
- Y. Konuma and T. Asakura. Effects of microphone mounting location and gender on accuracy in speech recognition using a throat microphone. *JASA Express Letters*, 3(9):095203, 2023. ISSN 2691-1191. doi: 10.1121/10.0020988. URL <https://doi.org/10.1121/10.0020988>.
- Suraj Kothawade, Anmol Mekala, D.Chandra Sekhara Hetha Havya, Mayank Kothiyari, Rishabh Iyer, Ganesh Ramakrishnan, and Preethi Jyothi. DITTO: Data-efficient and Fair Targeted Subset Selection for ASR Accent Adaptation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5810–5822. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-long.319. URL <https://doi.org/10.18653/v1/2023.acl-long.319>.

-
- Alkis Koudounas, Flavio Giobergia, Eliana Pastor, and Elena Baralis. A Contrastive Learning Approach to Mitigate Bias in Speech Models. In *Interspeech 2024*, pp. 827–831. ISCA, 2024a. doi: 10.21437/interspeech.2024-1219. URL <https://doi.org/10.21437/interspeech.2024-1219>.
- Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Luca de Alfaro, and Elena Baralis. Prioritizing Data Acquisition for end-to-end Speech Model Improvement. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7000–7004. IEEE, 2024b. doi: 10.1109/icassp48485.2024.10446326. URL <https://doi.org/10.1109/icassp48485.2024.10446326>.
- Alkis Koudounas, Eliana Pastor, Luca de Alfaro, and Elena Baralis. Mitigating Subgroup Disparities in Speech Models: A Divergence-Aware Dual Strategy. *IEEE Transactions on Audio, Speech and Language Processing*, 33:883–895, 2025a. ISSN 2998-4173. doi: 10.1109/taslpro.2025.3539429. URL <https://doi.org/10.1109/taslpro.2025.3539429>.
- Alkis Koudounas, Eliana Pastor, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Elisa Reale, Luca Cagliero, Sandro Cumani, Luca De Alfaro, Elena Baralis, and Daniele Amberti. Privacy Preserving Data Selection for Bias Mitigation in Speech Models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 6: Industry Track)*, pp. 738–748. Association for Computational Linguistics, 2025b. ISBN 979-8-89176-288-6. doi: 10.18653/v1/2025.acl-industry.52. URL <https://doi.org/10.18653/v1/2025.acl-industry.52>.
- Aravind Krishnan, Badr M. Abdullah, and Dietrich Klakow. On the Encoding of Gender in Transformer-based ASR Representations. In *Interspeech 2024*, pp. 3090–3094. ISCA, 2024. doi: 10.21437/interspeech.2024-2209. URL <https://doi.org/10.21437/interspeech.2024-2209>.
- Chun-Yi Kuan and Hung-yi Lee. Gender Bias in Instruction-Guided Speech Synthesis Models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 5387–5413. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.298. URL <https://doi.org/10.18653/v1/2025.findings-naacl.298>.
- Marek Kubis, Paweł Skórzewski, Iwona Christop, Mateusz Czyżnikiewicz, Jakub Kubiak, Łukasz Bondaruk, and Marcin Lewandowski. Preservation of Language Understanding Capabilities in Speech-aware Large Language Models, 2025. URL <https://arxiv.org/abs/2509.12171>.
- Roland Kuhn, Fineen Davis, Alain Désilets, Eric Joanis, Anna Kazantseva, Rebecca Knowles, Patrick Littell, Delaney Lothian, Aidan Pine, Caroline Running Wolf, Eddie Santos, Darlene Stewart, Gilles Boulianne, Vishwa Gupta, Brian Maracle Owennatékhá, Akwiratékhá’ Martin, Christopher Cox, Marie-Odile Junker, Olivia Sammons, Delasie Torkornoo, Nathan Thanyehténhas Brinklow, Sara Child, Benoît Farley, David Huggins-Daines, Daisy Rosenblum, and Heather Souter. The Indigenous Languages Technology project at NRC Canada: An empowerment-oriented approach to developing language software. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5866–5878. International Committee on Computational Linguistics, 2020. doi: 10.18653/v1/2020.coling-main.516. URL <https://doi.org/10.18653/v1/2020.coling-main.516>.
- Tahu Kukutai and John Taylor. *Indigenous data sovereignty: Toward an agenda*, volume 38. ANU Press, 2016. ISBN 9781760460303. doi: 10.22459/CAEPR38.11.2016. URL <https://press.anu.edu.au/publications/series/caepr/indigenous-data-sovereignty>.
- Matt J. Kusner, Joshua R. Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness, 2017. URL <https://doi.org/10.48550/arxiv.1703.06856>.
- Alexander Kwako, Yixin Wan, Jieyu Zhao, Mark Hansen, Kai-Wei Chang, and Li Cai. Does BERT Exacerbate Gender or L1 Biases in Automated English Speaking Assessment? In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pp. 668–681. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.bea-1.54. URL <https://doi.org/10.18653/v1/2023.bea-1.54>.

-
- Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. LLMs Get Lost In Multi-Turn Conversation, 2025. URL <https://arxiv.org/abs/2505.06120>.
- Rimita Lahiri, Manoj Kumar, Somer Bishop, and Shrikanth Narayanan. Learning Domain Invariant Representations for Child-Adult Classification from Speech. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6749–6753. IEEE, 2020. doi: 10.1109/icassp40776.2020.9054276. URL <https://doi.org/10.1109/icassp40776.2020.9054276>.
- Li-Fang Lai and Nicole Holliday. Exploring Sources of Racial Bias in Automatic Speech Recognition through the Lens of Rhythmic Variation. In *INTERSPEECH 2023*, pp. 1284–1288. ISCA, 2023. doi: 10.21437/interspeech.2023-159. URL <https://doi.org/10.21437/interspeech.2023-159>.
- Retno Larasati. Inclusivity of AI Speech in Healthcare: A Decade Look Back. In *Speech AI for All Workshop at CHI 2025*, 2025. URL <https://oro.open.ac.uk/105547/>.
- Colin Lea, Zifang Huang, Jaya Narain, Lauren Tooley, Dianna Yee, Dung Tien Tran, Panayiotis Georgiou, Jeffrey P Bigham, and Leah Findlater. From User Perceptions to Technical Improvement: Enabling People Who Stutter to Better Use Speech Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–16. ACM, 2023. doi: 10.1145/3544548.3581224. URL <https://doi.org/10.1145/3544548.3581224>.
- Ji-Na Lee and Ji-Yeoun Lee. An Efficient SMOTE-Based Deep Learning Model for Voice Pathology Detection. *Applied Sciences*, 13(6):3571, 2023. ISSN 2076-3417. doi: 10.3390/app13063571. URL <https://doi.org/10.3390/app13063571>.
- Tony Lee, Haoqin Tu, Chi Heem Wong, Zijun Wang, Siwei Yang, Yifan Mai, Yuyin Zhou, Cihang Xie, and Percy Liang. Ahelm: A holistic evaluation of audio-language models, 2025a. URL <https://doi.org/10.48550/arxiv.2508.21376>.
- Wonjun Lee, Solee Im, Heejin Do, Yunsu Kim, Jungseul Ok, and Gary Lee. DyPCL: Dynamic Phoneme-level Contrastive Learning for Dysarthric Speech Recognition. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4701–4712. Association for Computational Linguistics, 2025b. doi: 10.18653/v1/2025.naacl-long.240. URL <https://doi.org/10.18653/v1/2025.naacl-long.240>.
- Daniel Leix Palumbo and Robert Prey. Sounding out voice biometrics: Comparing and contrasting how the state and the private sector determine identity through voice. *Big Data & Society*, 11(4):20539517241297889, 2024. ISSN 2053-9517, 2053-9517. doi: 10.1177/20539517241297889. URL <https://doi.org/10.1177/20539517241297889>.
- Gianna Leoni, Lee Steven, Tūreiti Keith, Keoni Mahelona, Peter-Lucas Jones, and Suzanne Duncan. Solving Failure Modes in the Creation of Trustworthy Language Technologies. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 325–330. European Language Resources Association (ELRA) and ICCL, 2024. doi: 10.63317/4c2uqdp7r4s. URL <https://doi.org/10.63317/4c2uqdp7r4s>.
- Anna Leschanowsky and Sneha Das. Examining the interplay between privacy and fairness for speech processing: A review and perspective. In *4th Symposium on Security and Privacy in Speech Communication*, pp. 1–11. ISCA, 2024. doi: 10.21437/spsc.2024-1. URL <https://doi.org/10.21437/spsc.2024-1>.
- Lawrence Lessig. *Code and Other Laws of Cyberspace*. Basic Books, 1999. ISBN 9780465039128. URL <https://search.worldcat.org/title/43836713>.
- Wing-Zin Leung, Mattias Cross, Anton Ragni, and Stefan Goetze. Training Data Augmentation for Dysarthric Automatic Speech Recognition by Text-to-Dysarthric-Speech Synthesis. In *Interspeech 2024*, pp. 2494–2498. ISCA, 2024. doi: 10.21437/interspeech.2024-1645. URL <https://doi.org/10.21437/interspeech.2024-1645>.

-
- Changye Li, Trevor Cohen, and Serguei Pakhomov. Reexamining racial disparities in automatic speech recognition performance: The role of confounding by provenance, 2024. URL <https://doi.org/10.48550/arxiv.2407.13982>.
- Chin-Jou Li, Eunjung Yeo, Kwanghee Choi, Paula Andrea Pérez-Toro, Masao Someki, Rohan Kumar Das, Zhengjun Yue, Juan Rafael Orozco-Arroyave, Elmar Nöth, and David R. Mortensen. Towards Inclusive ASR: Investigating Voice Conversion for Dysarthric Speech Recognition in Low-Resource Languages. In *Interspeech 2025*, pp. 2128–2132. ISCA, 2025a. doi: 10.21437/interspeech.2025-512. URL <https://doi.org/10.21437/interspeech.2025-512>.
- Jinyu Li. Recent advances in end-to-end automatic speech recognition. *APSIPA Transactions on Signal and Information Processing*, 11(1):1–64, 2022. ISSN 2048-7703, 2048-7703. doi: 10.1561/116.00000050. URL <https://doi.org/10.1561/116.00000050>.
- Kai Li, Can Shen, Yile Liu, Jirui Han, Kelong Zheng, Xuechao Zou, Lionel Z. Wang, Shun Zhang, Xingjian Du, Hanjun Luo, Yingbin Jin, Xinxin Xing, Ziyang Ma, Yue Liu, Yifan Zhang, Junfeng Fang, Kun Wang, Yibo Yan, Gelei Deng, Haoyang Li, Yiming Li, Xiaobin Zhuang, Tianlong Chen, Qingsong Wen, Tianwei Zhang, Yang Liu, Haibo Hu, Zhizheng Wu, Xiaolin Hu, Eng-Siong Chng, Wenyuan Xu, XiaoFeng Wang, Wei Dong, and Xinfeng Li. AudioTrust: Benchmarking The Multifaceted Trustworthiness of Audio Large Language Models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 24938–25016, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/2aa212d6f40c1cb19b777e83db00ec6a-Paper-Conference.pdf.
- Qisheng Li and Shaomei Wu. Towards Fair and Inclusive Speech Recognition for Stuttering: Community-led Chinese Stuttered Speech Dataset Creation and Benchmarking. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pp. 1–9. ACM, 2024a. doi: 10.1145/3613905.3650950. URL <https://doi.org/10.1145/3613905.3650950>.
- Qisheng Li and Shaomei Wu. "I Want to Publicize My Stutter": Community-led Collection and Curation of Chinese Stuttered Speech Data. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW2): 1–27, 2024b. ISSN 2573-0142. doi: 10.1145/3687014. URL <https://doi.org/10.1145/3687014>.
- Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 986–995. Asian Federation of Natural Language Processing, 2017. URL <https://aclanthology.org/I17-1099/>.
- Yuxuan Li, Hirokazu Shirado, and Sauvik Das. Actions Speak Louder than Words: Agent Decisions Reveal Implicit Biases in Language Models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 3303–3325. ACM, 2025b. doi: 10.1145/3715275.3732212. URL <https://doi.org/10.1145/3715275.3732212>.
- Zhaolin Li and Jan Niehues. Multimodal In-context Learning for ASR of Low-resource Languages. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 24745–24760. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1239. URL <https://doi.org/10.18653/v1/2026.findings-acl.1239>.
- Siyu Liang, Nicolas Ballier, Gina-Anne Levow, and Richard Wright. Beyond WER: Probing Whisper’s Sub-token Decoder Across Diverse Language Resource Levels. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 31225–31235. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1591. URL <https://doi.org/10.18653/v1/2025.emnlp-main.1591>. arXiv:2509.25516.
- Siyu Liang, Nicolas Ballier, Gina-Anne Levow, and Richard Wright. The Limits of Data Scaling: Sub-token Utilization and Acoustic Saturation in Multilingual ASR. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 5785–5793. European Language Resources Association (ELRA), 2026. doi: 10.63317/5pht8xx8yosb. URL <https://doi.org/10.63317/5pht8xx8yosb>.

-
- Guan-Ting Lin, Shang-Wen Li, and Hung-yi Lee. Listen, Adapt, Better WER: Source-free Single-utterance Test-time Adaptation for Automatic Speech Recognition. In *Interspeech 2022*, pp. 2198–2202. ISCA, 2022. doi: 10.21437/interspeech.2022-600. URL <https://doi.org/10.21437/interspeech.2022-600>.
- Yi-Cheng Lin, Wei-Chih Chen, and Hung-Yi Lee. Spoken Stereoset: on Evaluating Social Bias Toward Speaker in Speech Large Language Models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 871–878. IEEE, 2024a. doi: 10.1109/slt61566.2024.10832259. URL <https://doi.org/10.1109/slt61566.2024.10832259>.
- Yi-Cheng Lin, Tzu-Quan Lin, Hsi-Che Lin, Andy T. Liu, and Hung-yi Lee. On the Social Bias of Speech Self-Supervised Models. In *Interspeech 2024*, pp. 4638–4642. ISCA, 2024b. doi: 10.21437/interspeech.2024-454. URL <https://doi.org/10.21437/interspeech.2024-454>.
- Yi-Cheng Lin, Tzu-Quan Lin, Chih-Kai Yang, Ke-Han Lu, Wei-Chih Chen, Chun-Yi Kuan, and Hung-Yi Lee. Listen and Speak Fairly: a Study on Semantic Gender Bias in Speech Integrated Large Language Models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pp. 439–446. IEEE, 2024c. doi: 10.1109/slt61566.2024.10832317. URL <https://doi.org/10.1109/slt61566.2024.10832317>.
- Yi-Cheng Lin, Haibin Wu, Huang-Cheng Chou, Chi-Chun Lee, and Hung-yi Lee. Emo-bias: A Large Scale Evaluation of Social Bias on Speech Emotion Recognition. In *Interspeech 2024*, pp. 4633–4637. ISCA, 2024d. doi: 10.21437/interspeech.2024-1073. URL <https://doi.org/10.21437/interspeech.2024-1073>.
- Yi-Cheng Lin, Jia-Hung Chen, and Hung-Yi Lee. MMMOS: Multi-domain Multi-axis Audio Quality Assessment. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–4. IEEE, 2025a. doi: 10.1109/asru65441.2025.11433844. URL <https://doi.org/10.1109/asru65441.2025.11433844>.
- Yi-Cheng Lin, Huang-Cheng Chou, and Hung-yi Lee. Mitigating Subgroup Disparities in Multi-Label Speech Emotion Recognition: A Pseudo-Labeling and Unsupervised Learning Approach. In *Interspeech 2025*, pp. 2053–2057. ISCA, 2025b. doi: 10.21437/interspeech.2025-662. URL <https://doi.org/10.21437/interspeech.2025-662>.
- Yi-Cheng Lin, Huang-Cheng Chou, Yu-Hsuan Li Liang, and Hung-Yi Lee. EMO-Debias: Benchmarking Gender Debiasing Techniques in Multi-Label Speech Emotion Recognition. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2025c. doi: 10.1109/asru65441.2025.11433837. URL <https://doi.org/10.1109/asru65441.2025.11433837>.
- Yi-Cheng Lin, Huang-Cheng Chou, Tzu-Chieh Wei, Kuan-Yu Chen, and Hung-Yi Lee. Do You Hear What I Mean? Quantifying the Instruction-Perception Gap in Instruction-Guided Expressive Text-To-Speech Systems. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 16472–16476. IEEE, 2026a. doi: 10.1109/icassp55912.2026.11462935. URL <https://doi.org/10.1109/icassp55912.2026.11462935>.
- Yi-Cheng Lin, Yusuke Hirota, Sung-Feng Huang, and Hung yi Lee. VIBE: Voice-Induced open-ended Bias Evaluation for Large Audio-Language Models via Real-World Speech, 2026b. URL <https://arxiv.org/abs/2604.17248>.
- Yi-Cheng Lin, Yu-Hsuan Li Liang, Hsuan Su, Tzu-Quan Lin, Shang-Tse Chen, Yun-Nung Chen, and Hung-yi Lee. Pseudo2Real: Task Arithmetic for Pseudo-Label Correction in Automatic Speech Recognition. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 1160–1175. Association for Computational Linguistics, 2026c. doi: 10.18653/v1/2026.findings-acl.59. URL <https://doi.org/10.18653/v1/2026.findings-acl.59>.
- Rosina Lippi-Green. *English with an accent: Language, ideology and discrimination in the United States*. Routledge, 2012. ISBN 9780415559119. URL <https://www.routledge.com/English-with-an-Accent-Language-Ideology-and-Discrimination-in-the-United-States/Lippi-Green/p/book/9780415559119>.

-
- Chunxi Liu, Michael Picheny, Leda Sari, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6162–6166. IEEE, 2022. doi: 10.1109/icassp43922.2022.9747501. URL <https://doi.org/10.1109/icassp43922.2022.9747501>.
- Evan Zheran Liu, Behzad Haghighi, Annie S. Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just Train Twice: Improving Group Robustness without Training Group Information, 2021a. URL <https://doi.org/10.48550/arxiv.2107.09044>.
- Hexin Liu, Haoyang Zhang, Qiquan Zhang, Xiangyu Zhang, Dongyuan Shi, Eng Siong Chng, and Haizhou Li. Code-Switching Speech Recognition Under the Lens: Model- and Data-Centric Perspectives. *IEEE Transactions on Audio, Speech and Language Processing*, 34:1853–1865, 2026. ISSN 2998-4173. doi: 10.1109/taslpro.2026.3675776. URL <https://doi.org/10.1109/taslpro.2026.3675776>.
- Jiashuo Liu, Zheyang Shen, Peng Cui, Linjun Zhou, Kun Kuang, Bo Li, and Yishi Lin. Stable adversarial learning under distributional shifts. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10): 8662–8670, 2021b. ISSN 2374-3468, 2159-5399. doi: 10.1609/aaai.v35i10.17050. URL <https://doi.org/10.1609/aaai.v35i10.17050>.
- Mingyi Liu. The Alignment Tax: Response Homogenization in Aligned LLMs and Its Implications for Uncertainty Estimation, 2026. URL <https://arxiv.org/abs/2603.24124>.
- Wei Liu, Ying Qin, Zhiyuan Peng, and Tan Lee. Sparsely Shared Lora on Whisper for Child Speech Recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 11751–11755. IEEE, 2024. doi: 10.1109/icassp48485.2024.10447004. URL <https://doi.org/10.1109/icassp48485.2024.10447004>.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL <https://doi.org/10.18653/v1/2023.emnlp-main.153>.
- Zhe Liu and Fuchun Peng. Statistical testing on ASR performance via blockwise bootstrap. In *Interspeech 2020*, pp. 596–600. ISCA, 2020. doi: 10.21437/interspeech.2020-1338. URL <https://doi.org/10.21437/interspeech.2020-1338>.
- Debbie Loakes. Does automatic speech recognition (ASR) have a role in the transcription of indistinct covert recordings for forensic purposes? *Frontiers in Communication*, 7:803452, 2022. ISSN 2297-900X. doi: 10.3389/fcomm.2022.803452. URL <https://doi.org/10.3389/fcomm.2022.803452>.
- Paola Lopez. Bias does not equal bias: A socio-technical typology of bias in data-based algorithmic systems. *Internet Policy Review*, 10(4):1–29, 2021. ISSN 2197-6775. doi: 10.14763/2021.4.1598. URL <https://doi.org/10.14763/2021.4.1598>.
- Anastassia Loukina, Nitin Madnani, and Klaus Zechner. The many dimensions of algorithmic fairness in educational applications. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pp. 1–10. Association for Computational Linguistics, 2019. doi: 10.18653/v1/w19-4401. URL <https://doi.org/10.18653/v1/w19-4401>.
- Daniel M. Low, Vishwanatha Rao, Gregory Randolph, Phillip C. Song, and Satrajit S. Ghosh. Identifying bias in models that detect vocal fold paralysis from audio recordings using explainable machine learning and clinician ratings. *PLOS Digital Health*, 3(5):e0000516, 2024. ISSN 2767-3170. doi: 10.1371/journal.pdig.0000516. URL <https://doi.org/10.1371/journal.pdig.0000516>.
- Fernando López, Paula Delgado-Santos, Pablo Gómez, David Solans, and Jordi Luque. "OK Aura, Be Fair With Me": Demographics-Agnostic Training for Bias Mitigation in Wake-up Word Detection, 2026. URL <https://arxiv.org/abs/2604.05830>.

-
- Anuprabha M, Krishna Gurugubelli, and Anil Kumar Vuppala. Fairness in Dysarthric Speech Synthesis: Understanding Intrinsic Bias in Dysarthric Speech Cloning using F5-TTS. In *Interspeech 2025*, pp. 2750–2754. ISCA, 2025. doi: 10.21437/interspeech.2025-1536. URL <https://doi.org/10.21437/interspeech.2025-1536>.
- Robert L. MacDonald, Pan-Pan Jiang, Julie Cattiau, Rus Heywood, Richard Cave, Katie Seaver, Marilyn A. Ladewig, Jimmy Tobin, Michael P. Brenner, Philip C. Nelson, Jordan R. Green, and Katrin Tomanek. Disordered Speech Data Collection: Lessons Learned at 1 Million Utterances from Project Euphonia. In *Interspeech 2021*, pp. 4833–4837. ISCA, 2021. doi: 10.21437/interspeech.2021-697. URL <https://doi.org/10.21437/interspeech.2021-697>.
- Samaneh Madanian, Talen Chen, Olayinka Adeleye, John Michael Templeton, Christian Poellabauer, Dave Parry, and Sandra L. Schneider. Speech emotion recognition using machine learning—A systematic review. *Intelligent Systems with Applications*, 20:200266, 2023. ISSN 2667-3053. doi: 10.1016/j.iswa.2023.200266. URL <https://doi.org/10.1016/j.iswa.2023.200266>.
- David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning Adversarially Fair and Transferable Representations. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 3381–3390. PMLR, 2018. URL <https://proceedings.mlr.press/v80/madras18a.html>.
- Amama Mahmood and Chien-Ming Huang. Gender Biases in Error Mitigation by Voice Assistants. *Proceedings of the ACM on Human-Computer Interaction*, 8(CSCW1):1–27, 2024. ISSN 2573-0142. doi: 10.1145/3637337. URL <https://doi.org/10.1145/3637337>.
- Lucas Maison and Yannick Estève. Some Voices are Too Common: Building Fair Speech Recognition Systems Using the CommonVoice Dataset. In *INTERSPEECH 2023*, pp. 4428–4432. ISCA, 2023. doi: 10.21437/interspeech.2023-2124. URL <https://doi.org/10.21437/interspeech.2023-2124>.
- H. B. Mann and D. R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947. ISSN 0003-4851. doi: 10.1214/aoms/1177730491. URL <https://doi.org/10.1214/aoms/1177730491>.
- Nina Markl. Language variation and algorithmic bias: understanding algorithmic bias in British English automatic speech recognition. In *2022 ACM Conference on Fairness Accountability and Transparency*, pp. 521–534. ACM, 2022. doi: 10.1145/3531146.3533117. URL <https://doi.org/10.1145/3531146.3533117>.
- Nina Markl and Stephen Joseph McNulty. Language technology practitioners as language managers: arbitrating data bias and predictive bias in ASR. In *Proceedings of the Language Resources and Evaluation Conference*, pp. 6328–6339. European Language Resources Association (ELRA), 2022. doi: 10.63317/4vo48eo9scth. URL <https://doi.org/10.63317/4vo48eo9scth>.
- Alicia Martin, Robert L. MacDonald, Pan-Pan Jiang, Marilyn Ladewig, Julie Cattiau, Rus Heywood, Richard Cave, Jimmy Tobin, Philip C. Nelson, and Katrin Tomanek. Project Euphonia: advancing inclusive speech recognition through expanded data collection and evaluation. *Frontiers in Language Sciences*, 4:1569448, 2025. ISSN 2813-4605. doi: 10.3389/flang.2025.1569448. URL <https://doi.org/10.3389/flang.2025.1569448>.
- Joshua L Martin. Spoken Corpora Data, Automatic Speech Recognition, and Bias Against African American Language: The case of Habitual’Be’. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 284–284. ACM, 2021. doi: 10.1145/3442188.3445893. URL <https://doi.org/10.1145/3442188.3445893>.
- Joshua L Martin and Kelly Elizabeth Wright. Bias in Automatic Speech Recognition: The Case of African American Language. *Applied Linguistics*, 44(4):613–630, 2023. ISSN 0142-6001, 1477-450X. doi: 10.1093/applin/amac066. URL <https://doi.org/10.1093/applin/amac066>.
- Natalia Martínez, Martín Bertrán, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *ICML*, pp. 6755–6764. PMLR, 2020. URL <http://proceedings.mlr.press/v119/martinez20a.html>.

-
- David McAllester and Karl Stratos. Formal Limitations on the Measurement of Mutual Information. In *International Conference on Artificial Intelligence and Statistics*, volume 108, pp. 875–884. PMLR, 2020. URL <https://proceedings.mlr.press/v108/mcallester20a.html>.
- Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6):1–35, 2022. ISSN 0360-0300, 1557-7341. doi: 10.1145/3457607. URL <https://doi.org/10.1145/3457607>.
- Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in gpt. In *Advances in Neural Information Processing Systems 35*, pp. 17359–17372. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022a. doi: 10.52202/068431-1262. URL <https://doi.org/10.52202/068431-1262>.
- Yen Meng, Yi-Hui Chou, Andy T. Liu, and Hung-yi Lee. Don’t Speak Too Fast: The Impact of Data Bias on Self-Supervised Speech Models. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3258–3262. IEEE, 2022b. doi: 10.1109/icassp43922.2022.9747897. URL <https://doi.org/10.1109/icassp43922.2022.9747897>.
- Zhong Meng, Jinyu Li, and Yifan Gong. Attentive Adversarial Learning for Domain-invariant Training. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6740–6744. IEEE, 2019. doi: 10.1109/icassp.2019.8683486. URL <https://doi.org/10.1109/icassp.2019.8683486>.
- Shira Michel, Sufi Kaur, Sarah Elizabeth Gillespie, Jeffrey Gleason, Christo Wilson, and Avijit Ghosh. “It’s not a representation of me”: Examining Accent Bias and Digital Exclusion in Synthetic AI Voice Services. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pp. 228–245. ACM, 2025. doi: 10.1145/3715275.3732018. URL <https://doi.org/10.1145/3715275.3732018>.
- N. Mirghafori, E. Fosler, and N. Morgan. Towards robustness to fast speech in ASR. In *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, volume 1, pp. 335–338. IEEE, 2002. doi: 10.1109/icassp.1996.541100. URL <https://doi.org/10.1109/icassp.1996.541100>.
- Vikramjit Mitra, Andreas Kathol, Jonathan D. Amith, and Rey Castillo García. Automatic Speech Transcription for Low-Resource Languages — The Case of YoloXóchitl Mixtec (Mexico). In *Interspeech 2016*, pp. 3076–3080. ISCA, 2016. doi: 10.21437/interspeech.2016-546. URL <https://doi.org/10.21437/interspeech.2016-546>.
- Vikramjit Mitra, Zifang Huang, Colin Lea, Lauren Tooley, Sarah Wu, Darren Botten, Ashwini Palekar, Shrinath Thelapurath, Panayiotis Georgiou, Sachin Kajarekar, and Jefferey Bigham. Analysis and Tuning of a Voice Assistant System for Dysfluent Speech. In *Interspeech 2021*, pp. 4848–4852. ISCA, 2021. doi: 10.21437/interspeech.2021-2006. URL <https://doi.org/10.21437/interspeech.2021-2006>.
- Aarish Shah Mohsin, Mohammad Nadeem, Shahab Saquib Sohail, Tughrul Arsalan, Mandar Gogate, Nasir Saleem, and Amir Hussain. Investigating Gender Bias in Text-to-Audio Generation Models. In *Interspeech 2025*, pp. 3369–3373. ISCA, 2025. doi: 10.21437/interspeech.2025-1488. URL <https://doi.org/10.21437/interspeech.2025-1488>.
- Ana Montalvo, Jose R. Calvo, and Jean-François Bonastre. Multi-Task Learning for Voice Related Recognition Tasks. In *Interspeech 2020*. ISCA, 2020. doi: 10.21437/interspeech.2020-1857. URL <https://doi.org/10.21437/interspeech.2020-1857>.
- Laureano Moro-Velazquez, JaeJin Cho, Shinji Watanabe, Mark A. Hasegawa-Johnson, Odette Scharenborg, Heejin Kim, and Najim Dehak. Study of the Performance of Automatic Speech Recognition Systems in Speakers with Parkinson’s Disease. In *Interspeech 2019*, pp. 3875–3879. ISCA, 2019. doi: 10.21437/interspeech.2019-2993. URL <https://doi.org/10.21437/interspeech.2019-2993>.
- Andrew Cameron Morris, Viktoria Maier, and Phil Green. From WER and RIL to MER and WIL: improved evaluation measures for connected speech recognition. In *Interspeech 2004*, pp. 2765–2768. ISCA, 2004. doi: 10.21437/interspeech.2004-668. URL <https://doi.org/10.21437/interspeech.2004-668>.

-
- Denis Moser, Nikola Stanic, and Murat Sariyar. Benchmarking speech-to-text robustness in noisy emergency medical dialogues: an evaluation of models under realistic acoustic conditions. *JAMIA Open*, 8(6):ooaf147, 2025. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooaf147. URL <https://doi.org/10.1093/jamiaopen/ooaf147>.
- Dena Mujtaba, Nihar Mahapatra, Megan Arney, J Yaruss, Hope Gerlach-Houck, Caryn Herring, and Jia Bin. Lost in Transcription: Identifying and Quantifying the Accuracy Biases of Automatic Speech Recognition Systems Against Disfluent Speech. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 4795–4809. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.naacl-long.269. URL <https://doi.org/10.18653/v1/2024.naacl-long.269>.
- Deirdre K. Mulligan, Joshua A. Kroll, Nitin Kohli, and Richmond Y. Wong. This Thing Called Fairness: Disciplinary Confusion Realizing a Value in Technology. *Proc. ACM Hum.-Comput. Interact.*, 2019. doi: 10.48550/arxiv.1909.11869. URL <https://doi.org/10.48550/arxiv.1909.11869>.
- Maya K. Nachesa and Vlad Niculae. kNN For Whisper And Its Effect On Bias And Speaker Adaptation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 6621–6627. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-naacl.369. URL <https://doi.org/10.18653/v1/2025.findings-naacl.369>.
- Iftekhar Naim, M. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. Automated prediction and analysis of job interview performance: The role of what you say and how you say it. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6. IEEE, 2015. doi: 10.1109/fg.2015.7163127. URL <https://doi.org/10.1109/fg.2015.7163127>.
- Iftekhar Naim, Md. Iftekhar Tanveer, Daniel Gildea, and Mohammed Ehsan Hoque. Automated analysis and prediction of job interview performance. *IEEE Transactions on Affective Computing*, 9(2):191–204, 2018. ISSN 1949-3045, 2371-9850. doi: 10.1109/taffc.2016.2614299. URL <https://doi.org/10.1109/taffc.2016.2614299>.
- Joyce Nakatumba-Nabende, Sulaiman Kagumire, Caroline Kantono, and Peter Nabende. A Systematic Literature Review on Bias Evaluation and Mitigation in Automatic Speech Recognition Models for Low-Resource African Languages. *ACM Computing Surveys*, 58(4):1–24, 2026. ISSN 0360-0300, 1557-7341. doi: 10.1145/3769089. URL <https://doi.org/10.1145/3769089>.
- Udhyakumar Nallasamy, Florian Metze, and Tanja Schultz. Active learning for accent adaptation in Automatic Speech Recognition. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pp. 360–365. IEEE, 2012. doi: 10.1109/slt.2012.6424250. URL <https://doi.org/10.1109/slt.2012.6424250>.
- Mikel K. Ngueajio and Gloria Washington. Hey ASR System! Why Aren’t You More Inclusive? In *Lecture Notes in Computer Science*, pp. 421–440. Springer Nature Switzerland, 2022. ISBN 9783031217067, 9783031217074. doi: 10.1007/978-3-031-21707-4_30. URL https://doi.org/10.1007/978-3-031-21707-4_30.
- Mewlude Nijat, Yang Wei, Shuailong Li, Abdusalam Dawut, and Askar Hamdulla. Beyond Native Norms: A Perceptually Grounded and Fair Framework for Automatic Speech Assessment. *Applied Sciences*, 16(2): 647, 2026. ISSN 2076-3417. doi: 10.3390/app16020647. URL <https://doi.org/10.3390/app16020647>.
- Helen Nissenbaum. Accountability in a computerized society. *Human values and the design of computer technology*, pp. 41–64, 1997.
- Paul-Gauthier Noe, Andreas Nautsch, Driss Matrouf, Pierre-Michel Bousquet, and Jean-Francois Bonastre. A Bridge between Features and Evidence for Binary Attribute-Driven Perfect Privacy. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3094–3098. IEEE, 2022. doi: 10.1109/icassp43922.2022.9746114. URL <https://doi.org/10.1109/icassp43922.2022.9746114>.

-
- Agnes Norbury, George Fairs, Alexandra L. Georgescu, Matthew M. Nour, Emilia Molimpakis, and Stefano Goria. A multimodal Bayesian network for symptom-level depression and anxiety prediction from voice and speech data. *Scientific Reports*, 16(1):5397, 2026. ISSN 2045-2322. doi: 10.1038/s41598-025-33331-w. URL <https://doi.org/10.1038/s41598-025-33331-w>.
- Esther Núñez-Vidal, Raúl Fernández-Ruiz, Agustín Álvarez Marquina, Irene Hidalgo-de-laGuía, Elena Garayzábal-Heinze, Nikola Hristov-Kalamov, Francisco Domínguez-Mateos, Cristina Conde, and Rafael Martínez-Olalla. Noninvasive Deep Learning Analysis for Smith–Magenis Syndrome Classification. *Applied Sciences*, 14(21):9747, 2024. ISSN 2076-3417. doi: 10.3390/app14219747. URL <https://doi.org/10.3390/app14219747>.
- Tobi Olatunji, Tejumade Afonja, Aditya Yadavalli, Chris Chinenye Emezue, Sahib Singh, Bonaventure F. P. Dossou, Joanne Osuchukwu, Salomey Osei, Atnafu Lambebo Tonja, Naome Etori, and Clinton Mbataku. AfriSpeech-200: Pan-African Accented Speech Dataset for Clinical and General Domain ASR. *Transactions of the Association for Computational Linguistics*, 11:1669–1685, 2023. ISSN 2307-387X. doi: 10.1162/tacl_a_00627. URL https://doi.org/10.1162/tacl_a_00627.
- Luca Oneto and Silvia Chiappa. Fairness in Machine Learning. In *Studies in Computational Intelligence*, pp. 155–196. Springer International Publishing, 2020. ISBN 9783030438821, 9783030438838. doi: 10.1007/978-3-030-43883-8_7. URL https://doi.org/10.1007/978-3-030-43883-8_7.
- Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*, volume 35, pp. 27730–27744. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2022. doi: 10.52202/068431-2011. URL <https://doi.org/10.52202/068431-2011>.
- Başak Ovacık. Digital Authority and the Reproduction of Gender Inequality: Addressing Gender Bias in Voice Assistant Development. *Journal of AI*, 9(1):13–31, 2025. doi: 10.61969/jai.1575060. URL <https://dergipark.org.tr/en/pub/jai/article/1575060>.
- Madhurananda Pahar, Caitlin Illingworth, Dorota Braun, Bahman Mirheidari, Lise Sproson, Daniel Blackburn, and Heidi Christensen. Can we trust AI to detect healthy multilingual English speakers among the cognitively impaired cohort in the UK? An investigation using real-world conversational speech, 2026. URL <https://arxiv.org/abs/2602.13047>.
- Ian Palmer, Andrew Rouditchenko, Andrei Barbu, Boris Katz, and James Glass. Spoken ObjectNet: A bias-controlled spoken caption dataset. In *Interspeech 2021*, pp. 3650–3654. ISCA, 2021. doi: 10.21437/interspeech.2021-245. URL <https://doi.org/10.21437/interspeech.2021-245>.
- Michele Panariello, Natalia Tomashenko, Xin Wang, Xiaoxiao Miao, Pierre Champion, Hubert Nourtel, Massimiliano Todisco, Nicholas Evans, Emmanuel Vincent, and Junichi Yamagishi. The VoicePrivacy 2022 Challenge: Progress and Perspectives in Voice Anonymisation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:3477–3491, 2024. ISSN 2329-9290, 2329-9304. doi: 10.1109/taslp.2024.3430530. URL <https://doi.org/10.1109/taslp.2024.3430530>.
- Zi Haur Pang, Xiaoxue Gao, Tatsuya Kawahara, and Nancy F. Chen. ERM-MinMaxGAP: Benchmarking and Mitigating Gender Bias in Multilingual Multimodal Speech-LLM Emotion Recognition. In *Interspeech 2026*, 2026. URL <https://arxiv.org/abs/2603.21050>. Accepted for publication.
- Orestis Papakyriakopoulos and Alice Xiang. Considerations for ethical speech recognition datasets. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, pp. 1287–1288. ACM, 2023. doi: 10.1145/3539597.3575793. URL <https://doi.org/10.1145/3539597.3575793>.
- Yeseul Park and Bowon Lee. Fine-tuning Strategies for Automatic Speech Recognition of Low-Resource Speech with Autism Spectrum Disorder. In *Interspeech 2025*, pp. 1858–1862. ISCA, 2025. doi: 10.21437/interspeech.2025-2535. URL <https://doi.org/10.21437/interspeech.2025-2535>.

-
- Ankita Pasad, Ju-Chieh Chou, and Karen Livescu. Layer-wise analysis of a self-supervised speech representation model. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 914–921. IEEE, 2021. doi: 10.1109/asru51503.2021.9688093. URL <https://doi.org/10.1109/asru51503.2021.9688093>.
- Tanvina Patel and Odette Scharenborg. Using Data Augmentations and VTLN to Reduce Bias in Dutch End-to-End Speech Recognition Systems, 2023. URL <https://arxiv.org/abs/2307.02009>.
- Tanvina Patel, Wiebke Hutiri, Aaron Yi Ding, and Odette Scharenborg. How to evaluate automatic speech recognition: Comparing different performance and bias measures, 2025. URL <https://arxiv.org/abs/2507.05885>.
- Amandalynne Paullada, Inioluwa Deborah Raji, Emily M. Bender, Emily Denton, and Alex Hanna. Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, 2(11):100336, 2021. ISSN 2666-3899. doi: 10.1016/j.patter.2021.100336. URL <https://doi.org/10.1016/j.patter.2021.100336>.
- Ameena L Payne, Tasha Austin, and Aris M Clemons. Beyond the Front Yard: The Dehumanizing Message of Accent-Altering Technology. *Applied Linguistics*, 45(3):553–560, 2024. ISSN 0142-6001, 1477-450X. doi: 10.1093/applin/amae002. URL <https://doi.org/10.1093/applin/amae002>.
- Thomas Pellegrini, Isabel Trancoso, Annika Hämäläinen, António Calado, Miguel Sales Dias, and Daniela Braga. Impact of Age in ASR for the Elderly: Preliminary Experiments in European Portuguese. In *Communications in Computer and Information Science*, pp. 139–147. Springer Berlin Heidelberg, 2012. ISBN 9783642352911, 9783642352928. doi: 10.1007/978-3-642-35292-8_15. URL https://doi.org/10.1007/978-3-642-35292-8_15.
- An-Ci Peng, Kuan-Tang Huang, Tien-Hong Lo, Hung-Shin Lee, Hsin-Min Wang, and Berlin Chen. Efficient Dialect-Aware Modeling and Conditioning for Low-Resource Taiwanese Hakka Speech Processing, 2026. URL <https://arxiv.org/abs/2602.22522>.
- Puyuan Peng, Brian Yan, Shinji Watanabe, and David Harwath. Prompting the Hidden Talent of Web-Scale Speech Models for Zero-Shot Task Generalization. In *INTERSPEECH 2023*, pp. 396–400. ISCA, 2023. doi: 10.21437/interspeech.2023-2032. URL <https://doi.org/10.21437/interspeech.2023-2032>.
- Raghuveer Peri, Krishna Somandepalli, and Shrikanth Narayanan. To train or not to train adversarially: A study of bias mitigation strategies for speaker recognition, 2022. URL <https://arxiv.org/abs/2203.09122>.
- Raghuveer Peri, Krishna Somandepalli, and Shrikanth Narayanan. A study of bias mitigation strategies for speaker recognition. *Computer Speech & Language*, 79:101481, 2023. ISSN 0885-2308. doi: 10.1016/j.csl.2022.101481. URL <https://doi.org/10.1016/j.csl.2022.101481>.
- Dana Pessach and Erez Shmueli. A Review on Fairness in Machine Learning. *ACM Computing Surveys*, 55(3): 1–44, 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3494672. URL <https://doi.org/10.1145/3494672>.
- Thao Phan. Amazon Echo and the aesthetics of whiteness. *Catalyst: Feminism, Theory, Technoscience*, 5(1): 1–38, 2019. ISSN 2380-3312, 2380-3312. doi: 10.28968/cftt.v5i1.29586. URL <https://doi.org/10.28968/cftt.v5i1.29586>.
- Peter Plantinga, Jen-Kai Chen, Roozbeh Sattari, Mirco Ravanelli, and Denise Klein. From Black Box to Biomarker: Sparse Autoencoders for Interpreting Speech Models of Parkinson’s Disease, 2025. URL <https://doi.org/10.48550/arxiv.2507.16836>.
- Crystal Min Hui Poon, Pai Chet Ng, Xiaoxiao Miao, Immanuel Jun Kai Loh, Bowen Zhang, Haoyu Song, and Ian McLoughlin. CLARITY: Contextual Linguistic Adaptation and Accent Retrieval for Dual-Bias Mitigation in Text-to-Speech Generation. In *2026 IEEE International Conference on Human-Machine Systems (ICHMS)*, pp. 548–553. IEEE, 2026. doi: 10.1109/ichms69701.2026.11602264. URL <https://doi.org/10.1109/ichms69701.2026.11602264>.

-
- A. Potamianos and S. Narayanan. Robust recognition of children’s speech. *IEEE Transactions on Speech and Audio Processing*, 11(6):603–616, 2003. ISSN 1063-6676. doi: 10.1109/tsa.2003.818026. URL <https://doi.org/10.1109/tsa.2003.818026>.
- Alisha Pradhan, Sheena Erete, Shaan Chopra, Pooja Upadhyay, Oluwaseun Sule, and Amanda Lazar. ‘No, not that voice again!’: Engaging Older Adults in Design of Anthropomorphic Voice Assistants. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–30, 2025. ISSN 2573-0142. doi: 10.1145/3711039. URL <https://doi.org/10.1145/3711039>.
- Kerri Prinos, Neal Patwari, and Cathleen A. Power. Speaking of accent: A content analysis of accent misconceptions in ASR research. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1245–1254. ACM, 2024. doi: 10.1145/3630106.3658969. URL <https://doi.org/10.1145/3630106.3658969>.
- Irina Proskurina, Guillaume Metzler, and Julien Velcin. The Other Side of Compression: Measuring Bias in Pruned Transformers. In *Lecture Notes in Computer Science*, pp. 366–378. Springer Nature Switzerland, 2023. ISBN 9783031300462, 9783031300479. doi: 10.1007/978-3-031-30047-9_29. URL https://doi.org/10.1007/978-3-031-30047-9_29.
- Dariia Puhach, Amir H. Payberah, and Éva Székely. Who Gets the Mic? Investigating Gender Bias in the Speaker Assignment of a Speech-LLM. In *Interspeech 2025*, pp. 2058–2062. ISCA, 2025. doi: 10.21437/interspeech.2025-1402. URL <https://doi.org/10.21437/interspeech.2025-1402>.
- Jinzi Qi and Hugo Van hamme. Parameter-efficient Dysarthric Speech Recognition Using Adapter Fusion and Householder Transformation. In *INTERSPEECH 2023*, pp. 151–155. ISCA, 2023. doi: 10.21437/interspeech.2023-1627. URL <https://doi.org/10.21437/interspeech.2023-1627>.
- Xiaoyi Qin, Na Li, Weng Chao, Dan Su, and Ming Li. Cross-Age Speaker Verification: Learning Age-Invariant Speaker Embeddings. In *Interspeech 2022*, pp. 1436–1440. ISCA, 2022. doi: 10.21437/interspeech.2022-648. URL <https://doi.org/10.21437/interspeech.2022-648>.
- Yangyang Qu, Massimiliano Todisco, Chiara Galdi, and Nicholas Evans. Fair-Gate: Fairness-Aware Interpretable Risk Gating for Sex-Fair Voice Biometrics. In *2026 14th International Workshop on Biometrics and Forensics (IWBF)*, pp. 1–6. IEEE, 2026. doi: 10.1109/iwbf68042.2026.11558146. URL <https://doi.org/10.1109/iwbf68042.2026.11558146>.
- Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. In *International Conference on Machine Learning*, volume 202, pp. 28492–28518. PMLR, 2023. URL <https://proceedings.mlr.press/v202/radford23a.html>.
- Rik Raes, Saskia E. Lensink, and Mykola Pechenizkiy. Everyone Deserves Their Voice to Be Heard: Analyzing Predictive Gender Bias in ASR Models Applied to Dutch Speech Data. In *Communications in Computer and Information Science*, pp. 19–34. Springer Nature Switzerland, 2026. ISBN 9783032253071, 9783032253088. doi: 10.1007/978-3-032-25308-8_2. URL https://doi.org/10.1007/978-3-032-25308-8_2.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36*, volume 36, pp. 53728–53741. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2023. doi: 10.52202/075280-2338. URL <https://doi.org/10.52202/075280-2338>.
- Anand Rai, Satyam Rahangdale, Utkarsh Anand, and Animesh Mukherjee. ASR-FAIRBENCH: Measuring and Benchmarking Equity Across Speech Recognition Systems, 2025. URL <https://doi.org/10.48550/arxiv.2505.11572>.
- Anand Kumar Rai, Siddharth D Jaiswal, Shubham Prakash, Bendi Pragnya Sree, and Animesh Mukherjee. DENOASR: Debiasing ASRs through Selective Denoising. In *2024 IEEE International Conference on*

-
- Knowledge Graph (ICKG)*, pp. 283–290. IEEE, 2024. doi: 10.1109/ickg63256.2024.00043. URL <https://doi.org/10.1109/ickg63256.2024.00043>.
- Desh Raj, David Snyder, Daniel Povey, and Sanjeev Khudanpur. Probing the Information Encoded in X-Vectors. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 726–733. IEEE, 2019. doi: 10.1109/asru46091.2019.9003979. URL <https://doi.org/10.1109/asru46091.2019.9003979>.
- Vishnu Raja, Adithya V Ganesan, Anand Syamkumar, Ritwik Banerjee, and H. Schwartz. Idiosyncratic Versus Normative Modeling of Atypical Speech Recognition: Dysarthric Case Studies. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 33514–33525. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1701. URL <https://doi.org/10.18653/v1/2025.emnlp-main.1701>.
- Jenalea Rajab, Anuoluwapo Aremu, Evelyn Asiko Chimoto, Dale Dunbar, Graham Morrissey, Fadel Thior, Luandrie Potgieter, Jessica Ojo, Atnafu Lambebo Tonja, Wilhelmina NdapewaOnyothi Nekoto, Pelonomi Moiloa, Jade Abbott, Vukosi Marivate, and Benjamin Rosman. The Esethu Framework: Reimagining Sustainable Dataset Governance and Curation for Low-Resource Languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30763–30776. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1487. URL <https://doi.org/10.18653/v1/2025.acl-long.1487>.
- Sai Sathiesh Rajan, Sakshi Udeshi, and Sudipta Chattopadhyay. AequiVox: Automated Fairness Testing of Speech Recognition Systems. In *Lecture Notes in Computer Science*, pp. 245–267. Springer International Publishing, 2022. ISBN 9783030994280, 9783030994297. doi: 10.1007/978-3-030-99429-7_14. URL https://doi.org/10.1007/978-3-030-99429-7_14.
- Sowmya Rasipuram, S.B. Pooja Rao, and Dinesh Babu Jayagopi. Automatic prediction of fluency in interface-based interviews. In *2016 IEEE Annual India Conference (INDICON)*, pp. 1–6. IEEE, 2016. doi: 10.1109/indicon.2016.7838991. URL <https://doi.org/10.1109/indicon.2016.7838991>.
- Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. Null It Out: Guarding Protected Attributes by Iterative Nullspace Projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 7237–7256. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.647. URL <https://doi.org/10.18653/v1/2020.acl-main.647>.
- Thomas Reitmaier, Electra Wallington, Dani Kalarikalayil Raju, Ondrej Klejch, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. Opportunities and Challenges of Automatic Speech Recognition Systems for Low-Resource Language Speakers. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–17. ACM, 2022. doi: 10.1145/3491102.3517639. URL <https://doi.org/10.1145/3491102.3517639>.
- Thomas Reitmaier, Electra Wallington, Ondrej Klejch, Dani Kalarikalayil Raju, Nina Markl, Emily Esther Nielsen, Gavin Bailey, Jennifer Pearson, Matt Jones, Peter Bell, and Simon Robinson. UnMute Toolkit: Speech Interactions Designed With Minoritised Language Speakers. In *ACM Conversational User Interfaces 2024*, pp. 1–6. ACM, 2024. doi: 10.1145/3640794.3665564. URL <https://doi.org/10.1145/3640794.3665564>.
- Wenze Ren, Yi-Cheng Lin, Wen-Chin Huang, Ryandhimas E. Zezario, Szu-Wei Fu, Sung-Feng Huang, Erica Cooper, Haibin Wu, Hung-Yu Wei, Hsin-Min Wang, Hung-Yi Lee, and Yu Tsao. HighRateMOS: Sampling-Rate Aware Modeling for Speech Quality Assessment. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–4. IEEE, 2025. doi: 10.1109/asru65441.2025.11434689. URL <https://doi.org/10.1109/asru65441.2025.11434689>.
- Wenze Ren, Yi-Cheng Lin, Wen-Chin Huang, Erica Cooper, Ryandhimas E. Zezario, Hsin-Min Wang, Hung-yi Lee, and Yu Tsao. MOS-Bias: From Hidden Gender Bias to Gender-Aware Speech Quality Assessment. In *Interspeech 2026*, 2026. URL <https://arxiv.org/abs/2603.10723>. Accepted for publication.

-
- Lauren Rhue. Racial influence on automated perceptions of emotions. *SSRN Electronic Journal*, 2018. ISSN 1556-5068. doi: 10.2139/ssrn.3281765. URL <https://doi.org/10.2139/ssrn.3281765>.
- Julius Richter, Yi-Chiao Wu, Steven Krenn, Simon Welker, Bunlong Lay, Shinji Watanabe, Alexander Richard, and Timo Gerkmann. EARS: An Anechoic Fullband Speech Dataset Benchmarked for Speech Enhancement and Dereverberation. In *Interspeech 2024*, pp. 4873–4877. ISCA, 2024. doi: 10.21437/interspeech.2024-153. URL <https://doi.org/10.21437/interspeech.2024-153>.
- Fabian Ritter-Gutierrez, Yi-Cheng Lin, Jui-Chiang Wei, Jeremy H.M. Wong, Nancy F. Chen, and Hung-Yi Lee. ASTAR-NTU solution to AudioMOS Challenge 2025 Track1. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–4. IEEE, 2025. doi: 10.1109/asru65441.2025.11434795. URL <https://doi.org/10.1109/asru65441.2025.11434795>.
- Morgane Riviere, Jade Copet, and Gabriel Synnaeve. ASR4REAL: An extended benchmark for speech models, 2021. URL <https://arxiv.org/abs/2110.08583>.
- Davit Rizhinashvili, Abdallah Hussein Sham, and Gholamreza Anbarjafari. Gender Neutralisation for Unbiased Speech Synthesising. *Electronics*, 11(10):1594, 2022. ISSN 2079-9292. doi: 10.3390/electronics11101594. URL <https://doi.org/10.3390/electronics11101594>.
- Nathan Roll, Calbert Graham, Yuka Tatsumi, Kim Tien Nguyen, Meghan Sumner, and Dan Jurafsky. In-Context Learning Boosts Speech Recognition via Human-like Adaptation to Speakers and Language Varieties. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 4412–4426. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.219. URL <https://doi.org/10.18653/v1/2025.emnlp-main.219>.
- A.E. Rosenberg. Automatic speaker verification: A review. *Proceedings of the IEEE*, 64(4):475–487, 1976. ISSN 0018-9219. doi: 10.1109/proc.1976.10156. URL <https://doi.org/10.1109/proc.1976.10156>.
- Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: 10.1016/0377-0427(87)90125-7. URL [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7).
- Neville Ryant, Kenneth Church, Christopher Cieri, Alejandrina Cristia, Jun Du, Sriram Ganapathy, and Mark Liberman. The Second DIHARD Diarization Challenge: Dataset, Task, and Baselines. In *Interspeech 2019*, pp. 978–982. ISCA, 2019. doi: 10.21437/interspeech.2019-1268. URL <https://doi.org/10.21437/interspeech.2019-1268>.
- Suhasini S and Bharathi B. ASR TAMIL SSN@ LT-EDI-2024: Automatic Speech Recognition system for Elderly People. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pp. 294–298. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.ltedi-1.40. URL <https://aclanthology.org/2024.ltedi-1.40/>.
- Pratik S. Sachdeva, Renata Barreto, Claudia von Vacano, and Chris J. Kennedy. Assessing Annotator Identity Sensitivity via Item Response Theory: A Case Study in a Hate Speech Corpus. In *2022 ACM Conference on Fairness Accountability and Transparency*, pp. 1585–1603. ACM, 2022. doi: 10.1145/3531146.3533216. URL <https://doi.org/10.1145/3531146.3533216>.
- Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally Robust Neural Networks. In *International Conference on Learning Representations*, 2020. URL https://iclr.cc/virtual_2020/poster_ryxGuJrFvS.html.
- Nithya Sambasivan, Erin Arnesen, Ben Hutchinson, Tulsee Doshi, and Vinodkumar Prabhakaran. Re-imagining Algorithmic Fairness in India and Beyond. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 315–328. ACM, 2021. doi: 10.1145/3442188.3445896. URL <https://doi.org/10.1145/3442188.3445896>.

-
- Mohammad Hossein Sameti, Sepehr Harfi Moridani, Ali Zarean, and Hossein Sameti. Accent-Invariant Automatic Speech Recognition via Saliency-Driven Spectrogram Masking. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 18487–18491. IEEE, 2026. doi: 10.1109/icassp55912.2026.11461358. URL <https://doi.org/10.1109/icassp55912.2026.11461358>.
- Leda Sar, Mark Hasegawa-Johnson, and Chang D. Yoo. Counterfactually Fair Automatic Speech Recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3515–3525, 2021. ISSN 2329-9290, 2329-9304. doi: 10.1109/taslp.2021.3126949. URL <https://doi.org/10.1109/taslp.2021.3126949>.
- Shree Harsha Bokkiahalli Satish, Gustav Eje Henter, and Éva Székely. Do Bias Benchmarks Generalise? Evidence from Voice-based Evaluation of Gender Bias in SpeechLLMs. In *2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, IEEE, 2026a. URL https://cmsworkshops.com/ICASSP2026/view_paper.php?PaperNum=14007.
- Shree Harsha Bokkiahalli Satish, Christoph Minixhofer, Maria Teleki, James Caverlee, Ondřej Klejch, Peter Bell, Gustav Eje Henter, and Éva Székely. The Voice Behind the Words: Quantifying Intersectional Bias in SpeechLLMs. In *Interspeech 2026*, 2026b. URL <https://arxiv.org/abs/2603.16941>. Accepted for publication.
- Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68. ACM, 2019a. doi: 10.1145/3287560.3287598. URL <https://doi.org/10.1145/3287560.3287598>.
- Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 59–68. ACM, 2019b. doi: 10.1145/3287560.3287598. URL <https://doi.org/10.1145/3287560.3287598>.
- Dana Serditova and Kevin Tang. A Sociolinguistic Analysis of Automatic Speech Recognition Bias in Newcastle English, 2026. URL <https://arxiv.org/abs/2603.24549>.
- Muhammad A. Shah, David Solans Noguero, Mikko A. Heikkilä, Bhiksha Raj, and Nicolas Kourtellis. Speech Robust Bench: A Robustness Benchmark For Speech Recognition. In *ICLR*, 2025. URL <https://openreview.net/forum?id=D0LuQNZfE1>.
- S. Shahnawazuddin, Nagaraj Adiga, Kunal Kumar, Aayushi Poddar, and Waquar Ahmad. Voice Conversion Based Data Augmentation to Improve Children’s Speech Recognition in Limited Data Scenario. In *Interspeech 2020*, pp. 4382–4386. ISCA, 2020. doi: 10.21437/interspeech.2020-1112. URL <https://doi.org/10.21437/interspeech.2020-1112>.
- Divya Sharma. EcoSpeak: Cost-Efficient Bias Mitigation for Partially Cross-Lingual Speaker Verification. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 379–394. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.27. URL <https://doi.org/10.18653/v1/2024.findings-naacl.27>.
- Tanusree Sharma, Yihao Zhou, and Visar Berisha. PRAC3 (Privacy, Reputation, Accountability, Consent, Credit, Compensation): Long Tailed Risks of Voice Actors in AI Data-Economy, 2025. URL <https://doi.org/10.48550/arxiv.2507.16247>.
- Renee Shelby, Shalaleh Rismani, Kathryn Henne, AJung Moon, Negar Rostamzadeh, Paul Nicholas, N’Mah Yilla-Akbari, Jess Gallegos, Andrew Smart, Emilio Garcia, and Gurleen Virk. Sociotechnical Harms of Algorithmic Systems: Scoping a Taxonomy for Harm Reduction. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 723–741. ACM, 2023. doi: 10.1145/3600211.3604673. URL <https://doi.org/10.1145/3600211.3604673>.
- Anushka Sanjay Shelke, Aditya Sneh, Arya Adyasha, and Haroon R. Lone. Fairness-Aware Few-Shot Learning for Audio-Visual Stress Detection, 2025. URL <https://doi.org/10.48550/arxiv.2511.09039>.

-
- Hua Shen, Yuguang Yang, Guoli Sun, Ryan Langman, Eunjung Han, Jasha Droppo, and Andreas Stolcke. Improving Fairness in Speaker Verification via Group-Adapted Fusion Network. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7077–7081. IEEE, 2022. doi: 10.1109/icassp43922.2022.9747384. URL <https://doi.org/10.1109/icassp43922.2022.9747384>.
- Zhonghao Shi, Xuan Shi, Anfeng Xu, Tiantian Feng, Harshvardhan Srivastava, Shrikanth Narayanan, and Maja Mataric. Examining Test-Time Adaptation for Personalized Child Speech Recognition. In *Interspeech 2025*, pp. 2820–2824. ISCA, 2025. doi: 10.21437/interspeech.2025-2513. URL <https://doi.org/10.21437/interspeech.2025-2513>.
- Vishal Shrivastava, Marisha Speights, and Akangkshya Pathak. Inclusive automatic speech recognition: A framework for equitable speech recognition in children with disorders. *The Journal of the Acoustical Society of America*, 157(4_Supplement):A24–A24, 2025. ISSN 1520-8524. doi: 10.1121/10.0037269. URL <https://doi.org/10.1121/10.0037269>.
- Sophie Si, Zhengxiong Li, and Wenyao Xu. Exploring Demographic Effects on Speaker Verification. In *2021 IEEE Conference on Communications and Network Security (CNS)*, pp. 1–2. IEEE, 2021. doi: 10.1109/cns53000.2021.9729038. URL <https://doi.org/10.1109/cns53000.2021.9729038>.
- Atli Sigurgeirsson and Eddie L. Ungless. Just Because We Camp, Doesn’t Mean We Should: The Ethics of Modelling Queer Voices. In *Interspeech 2024*, pp. 3050–3054. ISCA, 2024. doi: 10.21437/interspeech.2024-1982. URL <https://doi.org/10.21437/interspeech.2024-1982>.
- Anna Silnova, Themis Stafylakis, Ladislav Mošner, Oldřich Plchot, Johan Rohdin, Pavel Matějka, Lukáš Burget, Ondřej Glembek, and Niko Brummer. Analyzing Speaker Verification Embedding Extractors and Back-Ends Under Language and Channel Mismatch. In *The Speaker and Language Recognition Workshop (Odyssey 2022)*. ISCA, 2022. doi: 10.21437/odyssey.2022-2. URL <https://doi.org/10.21437/odyssey.2022-2>.
- Dipesh K. Singh, Preet P. Amin, Hardik B. Sailor, and Hemant A. Patil. Data Augmentation Using CycleGAN for End-to-End Children ASR. In *2021 29th European Signal Processing Conference (EUSIPCO)*, pp. 511–515. IEEE, 2021. doi: 10.23919/eusipco54536.2021.9616228. URL <https://doi.org/10.23919/eusipco54536.2021.9616228>.
- Rita Singh. *Profiling humans from their voice*, volume 41. Springer Singapore, 2019. ISBN 9789811384028, 9789811384035. doi: 10.1007/978-981-13-8403-5. URL <https://doi.org/10.1007/978-981-13-8403-5>.
- Amit Kumar Singh Yadav, Kratika Bhagatani, Paolo Bestagini, and Edward J. Delp. Improving Fairness in Synthetic Speech Detectors. In *2024 58th Asilomar Conference on Signals, Systems, and Computers*, pp. 362–366. IEEE, 2024a. doi: 10.1109/ieeconf60004.2024.10942853. URL <https://doi.org/10.1109/ieeconf60004.2024.10942853>.
- Amit Kumar Singh Yadav, Kratika Bhagatani, Davide Salvi, Paolo Bestagini, and Edward J. Delp. FairSSD: Understanding Bias in Synthetic Speech Detectors. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 4418–4428. IEEE, 2024b. doi: 10.1109/cvprw63382.2024.00445. URL <https://doi.org/10.1109/cvprw63382.2024.00445>.
- Isaac Slaughter, Craig Greenberg, Reva Schwartz, and Aylin Caliskan. Pre-trained Speech Processing Models Contain Human-Like Biases that Propagate to Speech Emotion Recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 8967–8989. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.findings-emnlp.602. URL <https://doi.org/10.18653/v1/2023.findings-emnlp.602>.
- Alex Solomonoff, Carl Quillen, and William M. Campbell. Channel compensation for SVM speaker recognition. In *Odyssey*, pp. 57–62, 2004. URL https://www.isca-archive.org/odyssey_2004/solomonoff04_odyssey.html.

-
- Rosy Southwell, Wayne Ward, Viet Anh Trinh, Charis Clevenger, Clay Clevenger, Emily Watts, Jason Reitman, Sidney D’Mello, and Jacob Whitehill. Automatic Speech Recognition Tuned for Child Speech in the Classroom. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12291–12295. IEEE, 2024. doi: 10.1109/icassp48485.2024.10447428. URL <https://doi.org/10.1109/icassp48485.2024.10447428>.
- Charan Sridhar and Shaomei Wu. J-j-j-just Stutter: Benchmarking Whisper’s Performance Disparities on Different Stuttering Patterns. In *Interspeech 2025*, pp. 3753–3757. ISCA, 2025. doi: 10.21437/interspeech.2025-2700. URL <https://doi.org/10.21437/interspeech.2025-2700>.
- Vojtěch Staněk, Karel Srna, Anton Firc, and Kamil Malinka. Scdf: A speaker characteristics deepfake speech dataset for bias analysis, 2025. URL <https://arxiv.org/abs/2508.07944>.
- Yolande Strengers and Jenny Kennedy. *The smart wife: Why Siri, Alexa, and other smart home devices need a feminist reboot*. The MIT Press, 2020. ISBN 9780262360036. doi: 10.7551/mitpress/12482.001.0001. URL <https://doi.org/10.7551/mitpress/12482.001.0001>.
- Hsuan Su, Hua Farn, Fan-Yun Sun, Shang-Tse Chen, and Hung-yi Lee. Task Arithmetic can Mitigate Synthetic-to-Real Gap in Automatic Speech Recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8905–8915. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.503. URL <https://doi.org/10.18653/v1/2024.emnlp-main.503>.
- Jinuo Sun, Yang Xiao, Sung Kyun Chung, Qiuchi Hu, Gongping Huang, Eun-Jung Holden, and Ting Dang. Activation Steering for Accent Adaptation in Large Audio Language Models. In *Interspeech 2026*, 2026. URL <https://arxiv.org/abs/2603.05813>. Accepted for publication.
- Sining Sun, Ching-Feng Yeh, Mei-Yuh Hwang, Mari Ostendorf, and Lei Xie. Domain Adversarial Training for Accented Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4854–4858. IEEE, 2018. doi: 10.1109/icassp.2018.8462663. URL <https://doi.org/10.1109/icassp.2018.8462663>.
- Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 1630–1640. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1159. URL <https://doi.org/10.18653/v1/p19-1159>.
- Harini Suresh and John Gutttag. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. In *Equity and Access in Algorithms, Mechanisms, and Optimization*, pp. 1–9. ACM, 2021. doi: 10.1145/3465416.3483305. URL <https://doi.org/10.1145/3465416.3483305>.
- Monorama Swain, Anna Katrine Van Zee, and Anders Søgaard. On Mitigating Performance Disparities in Multilingual Speech Recognition. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 5647–5655. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.323. URL <https://doi.org/10.18653/v1/2024.emnlp-main.323>.
- Monorama Swain, Bubai Maji, Jagabandhu Mishra, Markus Schedl, Anders Søgaard, and Jesper Rindom Jensen. Towards Fair ASR For Second Language Speakers Using Fairness Prompted Finetuning. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 17087–17091. IEEE, 2026. doi: 10.1109/icassp55912.2026.11464390. URL <https://doi.org/10.1109/icassp55912.2026.11464390>.
- Éva Székely, Joakim Gustafson, and Ilaria Torre. Prosody-controllable Gender-ambiguous Speech Synthesis: A Tool for Investigating Implicit Bias in Speech Perception. In *INTERSPEECH 2023*, pp. 1234–1238. ISCA, 2023. doi: 10.21437/interspeech.2023-2086. URL <https://doi.org/10.21437/interspeech.2023-2086>.
- Zeyu Tang, Jiji Zhang, and Kun Zhang. What-is and How-to for Fairness in Machine Learning: A Survey, Reflection, and Perspective. *ACM Computing Surveys*, 55(13s):1–37, 2023. ISSN 0360-0300, 1557-7341. doi: 10.1145/3597199. URL <https://doi.org/10.1145/3597199>.

-
- Srinath Tankasala, Long Chen, Andreas Stolcke, Anirudh Raju, Qianli Deng, Chander Chandak, Aparna Khare, Roland Maas, and Venkatesh Ravichandran. Cross-Utterance ASR Rescoring with Graph-Based Label Propagation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023. doi: 10.1109/icassp49357.2023.10096820. URL <https://doi.org/10.1109/icassp49357.2023.10096820>.
- Rachael Tatman. Gender and dialect bias in YouTube’s automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pp. 53–59. Association for Computational Linguistics, 2017. doi: 10.18653/v1/w17-1606. URL <https://doi.org/10.18653/v1/w17-1606>.
- Mike H.M. Teodorescu, Ionut Taraboanta, Sorin Andrian, Angela C. Ghiorghe, Horia-Nicolai Teodorescu, and Mihaela Grigorie. Fairness Constraints in Speech-to-Text Applications – Clinical Effects of Salivation. In *2023 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 72–77. IEEE, 2023. doi: 10.1109/sped59241.2023.10314880. URL <https://doi.org/10.1109/sped59241.2023.10314880>.
- Yannis Tevissen, Jérôme Boudy, Gérard Chollet, and Frédéric Petitpont. Towards measuring and scoring speaker diarization fairness, 2023. URL <https://doi.org/10.48550/arxiv.2302.09991>.
- Jimmy Tobin, Phillip Nelson, Bob MacDonald, Rus Heywood, Richard Cave, Katie Seaver, Antoine Desjardins, Pan-Pan Jiang, and Jordan R. Green. Automatic Speech Recognition of Conversational Speech in Individuals With Disordered Speech. *Journal of Speech, Language, and Hearing Research*, 67(11):4176–4185, 2024. ISSN 1092-4388, 1558-9102. doi: 10.1044/2024_jslhr-24-00045. URL https://doi.org/10.1044/2024_jslhr-24-00045.
- Katrin Tomanek, Vicky Zayats, Dirk Padfield, Kara Vaillancourt, and Fadi Biadisy. Residual Adapters for Parameter-Efficient ASR Adaptation to Atypical and Accented Speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6751–6760. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.541. URL <https://doi.org/10.18653/v1/2021.emnlp-main.541>.
- Andreas Triantafyllopoulos and Björn Schuller. Enrolment-based personalisation for improving individual-level fairness in speech emotion recognition. In *Interspeech 2024*, pp. 3729–3733. ISCA, 2024. doi: 10.21437/interspeech.2024-98. URL <https://doi.org/10.21437/interspeech.2024-98>.
- Aditay Tripathi, Aanchan Mohan, Saket Anand, and Maneesh Singh. Adversarial Learning of Raw Speech Features for Domain Invariant Speech Recognition. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5959–5963. IEEE, 2018. doi: 10.1109/icassp.2018.8462452. URL <https://doi.org/10.1109/icassp.2018.8462452>.
- Kumud Tripathi, Raj Gothi, and Pankaj Wasnik. Enhancing Whisper’s Accuracy and Speed for Indian Languages through Prompt-Tuning and Tokenization. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025. doi: 10.1109/icassp49660.2025.10887595. URL <https://doi.org/10.1109/icassp49660.2025.10887595>.
- Yun-Shao Tsai, Yi-Cheng Lin, Huang-Cheng Chou, and Hung-Yi Lee. CO-VADA: A Confidence-Oriented Voice Augmentation Debiasing Approach for Fair Speech Emotion Recognition. In *2025 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2025. doi: 10.1109/asru65441.2025.11434684. URL <https://doi.org/10.1109/asru65441.2025.11434684>.
- Shreya G. Upadhyay, Woan-Shiuan Chien, and Chi-Chun Lee. Is It Still Fair? Investigating Gender Fairness in Cross-Corpus Speech Emotion Recognition. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025. doi: 10.1109/icassp49660.2025.10889760. URL <https://doi.org/10.1109/icassp49660.2025.10889760>.
- Pol van Rijn and Pauline Larrouy-Maestri. Modelling individual and cross-cultural variation in the mapping of emotions to speech prosody. *Nature Human Behaviour*, 7(3):386–396, 2023. ISSN 2397-3374. doi: 10.1038/s41562-022-01505-5. URL <https://doi.org/10.1038/s41562-022-01505-5>.

-
- Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):205395171774353, 2017. ISSN 2053-9517, 2053-9517. doi: 10.1177/2053951717743530. URL <https://doi.org/10.1177/2053951717743530>.
- Irina-Elena Veliche and Pascale Fung. Improving Fairness and Robustness in End-to-End Speech Recognition Through Unsupervised Clustering. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023. doi: 10.1109/icassp49357.2023.10096836. URL <https://doi.org/10.1109/icassp49357.2023.10096836>.
- Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*, pp. 1–7. ACM, 2018. doi: 10.1145/3194770.3194776. URL <https://doi.org/10.1145/3194770.3194776>.
- Ravichander Vipperla, Steve Renals, and Joe Frankel. Ageing voices: The effect of changes in voice parameters on ASR performance. *EURASIP Journal on Audio, Speech, and Music Processing*, 2010(1):1–10, 2010. ISSN 1687-4714, 1687-4722. doi: 10.1155/2010/525783. URL <https://doi.org/10.1155/2010/525783>.
- Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *ICLR*, 2021. URL <https://openreview.net/forum?id=uXl3bZLkr3c>.
- Helin Wang, Venkatesh Ravichandran, Milind Rao, Becky Lammers, Myra Sydnor, Nicholas Maragakis, Ankur A. Butala, Jayne Zhang, Lora Clawson, Victoria Chovaz, and Laureano Moro-Velazquez. Improving fairness for spoken language understanding in atypical speech with Text-to-Speech, 2023a. URL <https://doi.org/10.48550/arxiv.2311.10149>.
- Rong Wang and Kun Sun. TIMIT Speaker Profiling: A Comparison of Multi-task learning and Single-task learning Approaches, 2024. URL <https://doi.org/10.48550/arxiv.2404.12077>.
- Siyin Wang, Chao-Han Yang, Ji Wu, and Chao Zhang. Can Whisper Perform Speech-Based In-Context Learning? In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 13421–13425. IEEE, 2024a. doi: 10.1109/icassp48485.2024.10446502. URL <https://doi.org/10.1109/icassp48485.2024.10446502>.
- Tianzi Wang, Shoukang Hu, Jiajun Deng, Zengrui Jin, Mengzhe Geng, Yi Wang, Helen Meng, and Xunying Liu. Hyper-parameter Adaptation of Conformer ASR Systems for Elderly and Dysarthric Speech Recognition. In *INTERSPEECH 2023*, pp. 1733–1737. ISCA, 2023b. doi: 10.21437/interspeech.2023-1263. URL <https://doi.org/10.21437/interspeech.2023-1263>.
- Yicheng Wang, Xiaotian Han, Leisheng Yu, and Na Zou. Beyond Fairness: Age-Harmless Parkinson’s Detection via Voice, 2023c. URL <https://doi.org/10.48550/arxiv.2309.13292>.
- Yicheng Wang, Mark Cusick, Mohamed Laila, Kate Puech, Zhengping Ji, Xia Hu, Michael Wilson, Noah Spitzer-Williams, Bryan Wheeler, and Yasser Ibrahim. FairLENS: Assessing Fairness in Law Enforcement Speech Recognition. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 3815–3824. IEEE, 2024b. doi: 10.1109/bigdata62323.2024.10825119. URL <https://doi.org/10.1109/bigdata62323.2024.10825119>.
- Yinong Oliver Wang, Nivedha Sivakumar, Falaah Arif Khan, Katherine Metcalf, Adam Golinski, Natalie Mackraz, Barry-John Theobald, Luca Zappella, and Nicholas Apostoloff. Is Your Model Fairly Certain? Uncertainty-Aware Fairness Evaluation for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 64343–64366. PMLR, 2025. URL <https://proceedings.mlr.press/v267/wang25cp.html>.
- Sheng-Lun Wei, Yu-Ling Liao, Yen-Hua Chang, Hen-Hsen Huang, and Hsin-Hsi Chen. Bias in the Ear of the Listener: Assessing Sensitivity in Audio Language Models Across Linguistic, Demographic, and Positional Variations. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 1570–1589. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-eacl.80. URL <https://doi.org/10.18653/v1/2026.findings-eacl.80>.

-
- Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel. Ethical and social risks of harm from language models, 2021. URL <https://doi.org/10.48550/arxiv.2112.04359>.
- Kimi Wenzel, Nitya Devireddy, Cam Davison, and Geoff Kaufman. Can Voice Assistants Be Microaggressors? Cross-Race Psychological Responses to Failures of Automatic Speech Recognition. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pp. 1–14. ACM, 2023. doi: 10.1145/3544548.3581357. URL <https://doi.org/10.1145/3544548.3581357>.
- Mark West, Rebecca Kraut, and Han Ei Chew. I’d blush if I could: closing gender divides in digital skills through education. Policy paper, UNESCO, 2019. URL <https://unesdoc.unesco.org/ark:/48223/pf0000367416>.
- Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80, 1945. ISSN 0099-4987. doi: 10.2307/3001968. URL <https://doi.org/10.2307/3001968>.
- Langdon Winner. Do Artifacts Have Politics? *Daedalus*, 109(1):121–136, 1980. URL <https://www.amacad.org/daedalus/modern-technology-problem-or-opportunity>.
- Walt Wolfram and Natalie Schilling. *American English: dialects and variation*. Wiley-Blackwell, 2015. ISBN 9781118390221. URL <https://www.wiley.com/en-us/Sociology%2Bof%2BLanguage-c-SOC0>.
- Dominika Woszczyk, Stavros Petridis, and David Millard. Domain Adversarial Neural Networks for Dysarthric Speech Recognition. In *Interspeech 2020*, pp. 3875–3879. ISCA, 2020. doi: 10.21437/interspeech.2020-2845. URL <https://doi.org/10.21437/interspeech.2020-2845>.
- Haibin Wu, Huang-Cheng Chou, Kai-Wei Chang, Lucas Goncalves, Jiawei Du, Jyh-Shing Roger Jang, Chi-Chun Lee, and Hung-Yi Lee. EMO-SUPERB: An in-depth look at speech emotion recognition, 2024. URL <https://doi.org/10.48550/arxiv.2402.13018>.
- Yihao Wu, Tianrui Wang, Yizhou Peng, Yi-Wen Chao, Xuyi Zhuang, Xinsheng Wang, Shunshun Yin, and Ziyang Ma. Evaluating Bias in Spoken Dialogue LLMs for Real-World Decisions and Recommendations. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 2056–2060. IEEE, 2026. doi: 10.1109/icassp55912.2026.11461562. URL <https://doi.org/10.1109/icassp55912.2026.11461562>.
- Sujie Xiao, Ta Li, Ruipeng Hao, Duojie Renzeng, and Qingwei Zhao. Multi-task learning method for accent and speech recognition based on spike features. *Journal of Tsinghua University (Science and Technology)*, 65(7):1310–1319, 2025. doi: 10.16511/j.cnki.qhdxxb.2025.26.029. URL <https://www.sciopen.com/article/10.16511/j.cnki.qhdxxb.2025.26.029>.
- Guangxuan Xu and Qingyuan Hu. Can model compression improve nlp fairness, 2022. URL <https://doi.org/10.48550/arxiv.2201.08542>.
- Chih-Kai Yang, Neo S. Ho, and Hung-yi Lee. Towards Holistic Evaluation of Large Audio-Language Models: A Comprehensive Survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 10155–10181. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.514. URL <https://doi.org/10.18653/v1/2025.emnlp-main.514>.
- Michael Yang, Abd-Allah El-Attar, and Theodora Chaspari. Deconstructing demographic bias in speech-based machine learning models for digital health. *Frontiers in Digital Health*, 6:1351637, 2024. ISSN 2673-253X. doi: 10.3389/fdgth.2024.1351637. URL <https://doi.org/10.3389/fdgth.2024.1351637>.
- Mu Yang and John H. L. Hansen. Activation Steering for Accent-Neutralized Zero-Shot Text-To-Speech, 2026. URL <https://arxiv.org/abs/2603.05977>.

-
- Yanbing Yang, Hao Shi, Yuqin Lin, Meng Ge, Longbiao Wang, Qingzhi Hou, and Jianwu Dang. Adaptive Attention Network with Domain Adversarial Training for Multi-Accent Speech Recognition. In *2022 13th International Symposium on Chinese Spoken Language Processing (ISCSLP)*, pp. 6–10. IEEE, 2022a. doi: 10.1109/iscslp57327.2022.10038197. URL <https://doi.org/10.1109/iscslp57327.2022.10038197>.
- Zehui Yang, Yifan Chen, Lei Luo, Runyan Yang, Lingxuan Ye, Gaofeng Cheng, Ji Xu, Yaohui Jin, Qingqing Zhang, Pengyuan Zhang, Lei Xie, and Yonghong Yan. Open source magicdata-ramc: A rich annotated mandarin conversational (ramc) speech dataset. In *Interspeech 2022*, pp. 1736–1740. ISCA, 2022b. doi: 10.21437/interspeech.2022-729. URL <https://doi.org/10.21437/interspeech.2022-729>.
- Han-Gyol Yi, Rajka Smiljanic, and Bharath Chandrasekaran. The neural processing of foreign-accented speech and its relationship to listener bias. *Frontiers in Human Neuroscience*, 8, 2014. ISSN 1662-5161. doi: 10.3389/fnhum.2014.00768. URL <https://doi.org/10.3389/fnhum.2014.00768>.
- Nurmemet Yolwas and Weijing Meng. JSUM: A Multitask Learning Speech Recognition Model for Jointly Supervised and Unsupervised Learning. *Applied Sciences*, 13(9):5239, 2023. ISSN 2076-3417. doi: 10.3390/app13095239. URL <https://doi.org/10.3390/app13095239>.
- Dong Yu and Li Deng. *Automatic speech recognition*, volume 1. Springer London, 2015. ISBN 9781447157786, 9781447157793. doi: 10.1007/978-1-4471-5779-3. URL <https://doi.org/10.1007/978-1-4471-5779-3>.
- Jun Yu, Xinlong Hao, Zeyu Cui, Peng He, and Tongliang Liu. Boosting Fairness for Masked Face Recognition. In *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 1531–1540. IEEE, 2021. doi: 10.1109/iccvw54120.2021.00178. URL <https://doi.org/10.1109/iccvw54120.2021.00178>.
- Anna Zee, Marc Zee, and Anders Sogaard. Group Fairness in Multilingual Speech Recognition Models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2213–2226. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-naacl.143. URL <https://doi.org/10.18653/v1/2024.findings-naacl.143>.
- Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning Fair Representations. In *Proceedings of the 30th International Conference on Machine Learning*, pp. 325–333. PMLR, 2013. URL <http://proceedings.mlr.press/v28/zemel13.pdf>.
- Dong Zhang, Zhaowei Li, Shimin Li, Xin Zhang, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. SpeechAlign: Aligning Speech Generation to Human Preferences. In *Advances in Neural Information Processing Systems 37*, pp. 50343–50360. Neural Information Processing Systems Foundation, Inc. (NeurIPS), 2024a. doi: 10.52202/079017-1594. URL <https://doi.org/10.52202/079017-1594>.
- Guangyan Zhang, Kaitao Song, Xu Tan, Daxin Tan, Yuzi Yan, Yanqing Liu, Gang Wang, Wei Zhou, Tao Qin, Tan Lee, and Sheng Zhao. Mixed-Phoneme BERT: Improving BERT with Mixed Phoneme and Sup-Phoneme Representations for Text to Speech. In *Interspeech 2022*, pp. 456–460. ISCA, 2022a. doi: 10.21437/interspeech.2022-621. URL <https://doi.org/10.21437/interspeech.2022-621>.
- Jicheng Zhang, Yizhou Peng, Van Tung Pham, Haihua Xu, Hao Huang, and Eng Siong Chng. E2E-Based Multi-Task Learning Approach to Joint Speech and Accent Recognition. In *Interspeech 2021*. ISCA, 2021. doi: 10.21437/interspeech.2021-1495. URL <https://doi.org/10.21437/interspeech.2021-1495>.
- Tuo Zhang, Tiantian Feng, Samiul Alam, Sunwoo Lee, Mi Zhang, Shrikanth S. Narayanan, and Salman Avestimehr. Fedaudio: A federated learning benchmark for audio tasks. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2023a. doi: 10.1109/icassp49357.2023.10096500. URL <https://doi.org/10.1109/icassp49357.2023.10096500>.
- Xiaojun Zhang, Changwei Zhou, Xincheng Zhu, Zhi Tao, and Heming Zhao. Class-imbalanced voice pathology classification: Combining hybrid sampling with optimal two-factor random forests. *Applied Acoustics*, 190:108618, 2022b. ISSN 0003-682X. doi: 10.1016/j.apacoust.2021.108618. URL <https://doi.org/10.1016/j.apacoust.2021.108618>.

-
- Yang Zhang, Zhiqiang Lv, Haibin Wu, Shanshan Zhang, Pengfei Hu, Zhiyong Wu, Hung-yi Lee, and Helen Meng. MFA-Conformer: Multi-scale Feature Aggregation Conformer for Automatic Speaker Verification. In *Interspeech 2022*, pp. 306–310. ISCA, 2022c. doi: 10.21437/interspeech.2022-563. URL <https://doi.org/10.21437/interspeech.2022-563>.
- Yixuan Zhang, Yuanyuan Zhang, Tanvina Patel, and Odette Scharenborg. Comparing data augmentation and training techniques to reduce bias against non-native accents in hybrid speech recognition systems. In *1st Workshop on Speech for Social Good (S4SG)*, pp. 15–19. ISCA, 2022d. doi: 10.21437/s4sg.2022-4. URL <https://doi.org/10.21437/s4sg.2022-4>.
- Yuanyuan Zhang, Yixuan Zhang, Bence Halpern, Tanvina Patel, and Odette Scharenborg. Mitigating bias against non-native accents. In *Interspeech 2022*, pp. 3168–3172. ISCA, 2022e. doi: 10.21437/interspeech.2022-836. URL <https://doi.org/10.21437/interspeech.2022-836>.
- Yuanyuan Zhang, Aaricia Herygers, Tanvina Patel, Zhengjun Yue, and Odette Scharenborg. Exploring Data Augmentation in Bias Mitigation Against Non-Native-Accented Speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pp. 1–8. IEEE, 2023b. doi: 10.1109/asru57964.2023.10389756. URL <https://doi.org/10.1109/asru57964.2023.10389756>.
- Yuanyuan Zhang, Zhengjun Yue, Tanvina Patel, and Odette Scharenborg. Improving child speech recognition with augmented child-like speech. In *Interspeech 2024*, pp. 5183–5187. ISCA, 2024b. doi: 10.21437/interspeech.2024-485. URL <https://doi.org/10.21437/interspeech.2024-485>.
- Yuanyuan Zhang et al. Exploring data augmentation in bias mitigation against non-native-accented speech. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023c.
- Shuyang Zhao, Mittul Singh, Abraham Woubie, and Reima Karhila. Data augmentation for children ASR and child-adult speaker classification using voice conversion methods. In *INTERSPEECH 2023*, pp. 4593–4597. ISCA, 2023. doi: 10.21437/interspeech.2023-702. URL <https://doi.org/10.21437/interspeech.2023-702>.
- Yue Zhao, Jianjian Yue, Wei Song, Xiaona Xu, Xiali Li, Licheng Wu, and Qiang Ji. Tibetan Multi-Dialect Speech and Dialect Identity Recognition. *Computers, Materials & Continua*, 60(3):1223–1235, 2019. ISSN 1546-2226. doi: 10.32604/cmc.2019.05636. URL <https://doi.org/10.32604/cmc.2019.05636>.
- 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, Suzanne Bakken, Kathryn H Bowles, Ian Spens, Nicole Onorato, Sridevi Sridharan, and Margaret V McDonald. 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, 2024. ISSN 2574-2531. doi: 10.1093/jamiaopen/ooae130. URL <https://doi.org/10.1093/jamiaopen/ooae130>.
- Maryam Zolnoori, Hossein Azadmaleki, Yasaman Haghbin, Ali Zolnour, Mohammad Javad Momeni Nezhad, Sina Rashidi, Mehdi Naserian, Elyas Esmaeili, and Sepehr Karimi Arpanahi. National Institute on Aging PREPARE Challenge: Early Detection of Cognitive Impairment Using Speech – The SpeechCARE Solution, 2025. URL <https://arxiv.org/abs/2511.08132>.