GEORG-AUGUST-UNIVERSITÄT GÖTTINGEN IN PUBLICA COMMODA SEIT 1737

Emilio Boldt, Jachin Tekle Gemta (Universität Hamburg)

Comparing Gender Bias in Classical Speech Emotion Recognition and AudioLLMs on EmoDB and RAVDESS

Speech Emotion Recognition (SER) is increasingly deployed in settings where reliability has direct consequences for people, including call-centre routing, mental-health and telehealth screening, driver monitoring, and voice assistants. Here, a system that recognises emotions less accurately for some speaker groups than others can translate into unequal service or misjudged distress in a healthcare setting. Reliable and equitable emotion recognition is therefore a precondition for responsible deployment.

A well-documented source of such inequity is speaker gender. Prior work shows that gender-dependent modelling can improve SER (Vogt and André, 2006) and that male–female performance differences persist in modern systems (Lin et al., 2024). This evidence, however, concerns classical pipelines built on acoustic features or embeddings plus a dedicated classifier. A newer paradigm, Audio Large Language Models (AudioLLMs), performs emotion recognition through general audio-language reasoning rather than task-specific classification (Chu et al., 2023). Whether these models inherit, reduce, or reshape known gender biases remains largely unexplored; Pang et al. (2026) provide one of the few existing investigations. Our work grounds itself in this literature and addresses that gap. We ask which differences in gender bias can be observed between a classical SER pipeline and AudioLLMs, and how stable those differences are across prompting strategies and languages, aiming to learn not only whether AudioLLMs are biased, but the form their bias takes and the mechanisms it is entangled with, knowledge necessary for evaluating such systems for fairness before deployment. We compare a "modern-classical" pipeline, Whisper embeddings classified by logistic regression and a multilayer perceptron under five-fold cross-validation, following George and Ilyas (2024), against three AudioLLMs (MERaLiON-3-10B, Qwen2-Audio-7B, Audio Flamingo 3) used zero-shot. We evaluate on two gender-balanced datasets of acted speech: EmoDB (535 German recordings, 10 speakers, 7 emotions) and RAVDESS (1,440 English recordings, 24 speakers, 8 emotions). Each recording is paired with a task instruction under four prompt conditions (baseline, prosody-focused, definition-based "expert", and German-language); outputs are normalised to the dataset labels, with out-of-taxonomy predictions counted as incorrect. We subtract separate male and female confusion matrices into a per-cell gender-bias score, reported alongside accuracy and unweighted average recall. The classical pipeline produces small, stable gender effects, whereas the AudioLLMs show larger, emotion-specific disparities entangled with vocabulary collapse and prompt sensitivity — including cases where the direction of bias reverses with the prompt.

Our research has implications on the level of social justice. We understand social justice here as distributive fairness in algorithmic systems: when a technology mediates access to services or care, its benefits and harms should not depend on a speaker's group membership. Gender bias in SER is a concrete instance, since the same expression is read differently depending on whether the speaker is heard as male or female — a disparity that, at scale, reproduces existing inequalities. Our central finding sharpens this for AudioLLMs: if fairness shifts with prompt wording, an audit under one prompt may declare a system fair while it behaves inequitably under another. Socially just deployment therefore requires prompt-aware, group-wise fairness evaluation rather than one-off accuracy testing — which our gender-bias analysis offers one practical instrument toward.