Twists, Humps, and Pebbles: Multilingual Speech Recognition Models Exhibit Gender Performance Gaps

Giuseppe Attanasio[♡], **Beatrice Savoldi**[♠], **Dennis Fucci**[♢♠], **Dirk Hovy**[♣]

[♡] Instituto de Telecomunicações, Lisbon, Portugal
[♢] University of Trento, Trento, Italy
[♠] Fondazione Bruno Kessler, Trento, Italy
[♣] Bocconi University, Milan, Italy
giuseppe.attanasio@lx.it.pt

Abstract

Current automatic speech recognition (ASR) models are designed to be used across many languages and tasks without substantial changes. However, this broad language coverage hides performance gaps *within* languages, for example, across genders. Our study systematically evaluates the performance of two widely used multilingual ASR models on three datasets, encompassing 19 languages from eight language families and two speaking conditions. Our findings reveal clear gender disparities, with the advantaged group varying across languages and models. Surprisingly, those gaps are not explained by acoustic or lexical properties. However, probing internal model states reveals a correlation with gendered performance gap. That is, the easier it is to distinguish speaker gender in a language using probes, the more the gap reduces, favoring female speakers. Our results show that gender disparities persist even in state-of-the-art models. Our findings have implications for the improvement of multilingual ASR systems, underscoring the importance of accessibility to training data and nuanced evaluation to predict and mitigate gender gaps. We release all code and artifacts at https://github.com/g8a9/multilingual-asr-gender-gap.

1 Introduction

A new class of multi-task, multilingual neural networks (Radford et al., 2022; Communication et al., 2023; Chu et al., 2023) has recently pushed the boundaries of several speech-related tasks, including automatic speech recognition (ASR). As these models offer support to an increasing number of languages at no cost, they have found widespread adoption and have been integrated into applications for the general public, such as real-time voice transcription.[1] However, while the user base expands, one question has yet to be answered: *will usability be the same for everyone?* In this paper, we seek an answer to this question by studying how systems understand the voices of different genders and across multiple languages. These two axes of analysis are motivated by complementary interests.

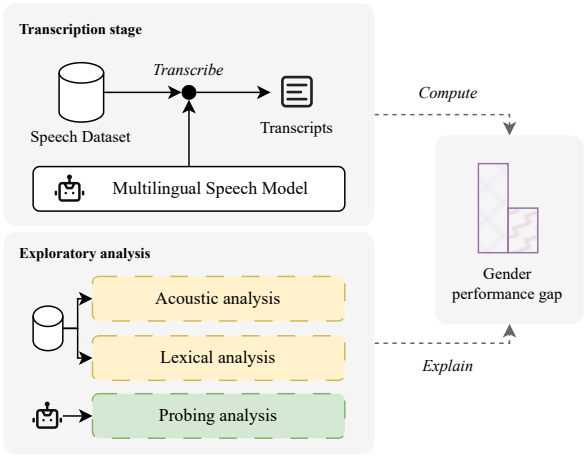


Figure 1: **Study overview.** We transcribe three speech datasets across 19 languages and compute gender performance gaps. Next, we investigate the models for possible causes of these gaps.

On the one hand, voice and language production are among the strongest identity traits. They vary across individuals, sociodemographic groups (Labov, 1964; Wolfram, 2004; Alim, 2004), and, crucially, across genders. Gendered differences in the voice—rooted in physiological factors linked to biological sex (e.g., vocal tract length), as well as in sociocultural ones (e.g., prescribed registers) (Zimman, 2017, 2020)—have been extensively studied in sociolinguistics (e.g., Coleman, 1976; Busby and Plant, 1995; Hillenbrand and Clark, 2009). Also,

[1] See, for instance, http://tcrn.ch/4el8Brr

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 21318–21340
November 12-16, 2024

empirical studies have shown that gender affects the performance of traditional ASR systems in English (Garnerin et al., 2019). On the other hand, voice and speech can change drastically across languages, cultures, *and* in function of gender. For example, males' and females' pitch—a measure of the highness of the voice—is much closer in Japanese- than in English-speaking people (Loveday, 1981; Yuasa, 2008).

Despite a large literature on gender, voice, and ASR, to our knowledge, no studies have tested whether multi-task, multilingual ASR models serve genders to a comparable level across languages. Arguably, overlooking potential disparities can result in unequal service quality for already socially disadvantaged individuals (Mengesha et al., 2021; Tatman, 2017).

To fill this gap, we systematically study gendered performance gaps in massively multilingual ASR models. We set out to answer two research questions. (**Q1**) *Do multi-task, multilingual ASR models perform equally well across speakers identifying as women, men, or neither of the two?* If so, (**Q2**), *can we relate gender gaps to acoustic and lexical variation in data or model internal states?*

To answer Q1, we evaluate two state-of-the-art multilingual open-weight ASR models on three datasets, covering 19 languages and two speech conditions (i.e. read and spontaneous). The results of our extensive evaluation show that **models systematically exhibit gender performance gaps**. However, whether models favor male or female speakers depends on the dataset and language. Results on one dataset also highlight subpar performance for speakers who do not identify with either gender compared to males.

While studying acoustic and lexical phenomena in test data (Q2), we found no significant correlations of specific features with performance gaps between male and female speakers. This objective finding underscores the complexity of the issue and the need for further research. Instead, interpretability analyses suggest that when presented with speech from men and women, models build different internal representations that can serve as a proxy for gender disparities.

Contributions. We conduct the first extensive evaluation of two widely used multilingual ASR models for gender performance differences. We document significant gender gaps by inspecting model internal states. We release code, data, and all artifacts we produce for future research.

Bias Statement. When using gender as a variable, we rely on speakers' declared identity (see §8). We evaluate whether speech from individuals identifying as female, male, or neither of the two is processed equally well by multilingual ASR models. We measure disparities of minority groups versus the socially advantaged group, i.e., men. Performance parity is the ideal outcome. We define a system to be biased if it risks further contributing allocative (i.e., quality of service, technology less accessible) and representational harms impacting the minority groups, e.g., feeding into stereotypes about the inadequacy of women and speech technology or "shrill," "incorrect" voice (Tallon, 2019).[2]

2 Background

Multi-task Multilingual Speech Models. Contemporary speech recognition systems process audio and text separately and use common cross-modal interactions—e.g., OpenAI's Whisper (Radford et al., 2022) is loosely inspired by the standard Transformer (Vaswani et al., 2017). Crucially, multitasking and multilinguality come at virtually no cost for the user. With Whisper, for example, special "task" and "language" tokens can be prepended to the decoder input to change its functions. This strategy allows multi-tasking and multilinguality without architectural changes or fine-tuning (see Figure 1, top-right, in Radford et al. (2022)). Similarly, Meta's SeamlessM4T (Communication et al., 2023) uses a speech encoder for audio and an encoder-decoder transformer for text (Costa-jussà et al., 2022, NLLB).

Gender and Speech (Technologies). Gendered aspects of the voice are one of the most salient individual traits (Kreiman and Sidtis, 2011; Azul, 2015; Zimman, 2021) and have long been studied in linguistics (Zimman, 2020). The anatomy and makeup of the vocal tract do play a role, as it determines pitch range and formant variations, often regarded as the most distinctive vocal features of cisgender men and women.[3] However, gender variation in the voice has been shown to arise from several aspects besides physical ones (Oates and

[2] As voice technology expert Tom Schalk once put it, "many issues with women's voices could be fixed if female drivers were willing to sit through lengthy training... Women could be taught to speak louder and direct their voices towards the microphone." https://bit.ly/time-shalck

[3] *Cisgender* describes individuals whose gender identity matches their birth-assigned sex (Fuchs and Toda, 2010).

Dacakis, 2015; Zimman, 2018; Becker et al., 2022). As studies including transgender individuals show (Zimman, 2017), sociocultural factors influence vocal use, including which parts of the pitch range are used (Loveday, 1981; Yuasa, 2008), and articulatory practices for sibilant consonants and vowels that are perceived as gender characterizing (Pharao et al., 2014; Podesva et al., 2016; Li, 2017).

For speech technologies, sociodemographic variation has posed challenges to ASR systems (Sawalha and Abu Shariah, 2013; Liu et al., 2022; Rajan et al., 2022; Fucci et al., 2023). Various works have found ASR models for English and French to recognize better male speech and voices (Tatman, 2017; Garnerin et al., 2019, 2021). This effect is often a result of the under-representation of women in the training data (Meyer et al., 2020; Gaido et al., 2020; Garnerin et al., 2020). However, various works have found the reverse to be true. Several studies found comparatively better performance for women in more spontaneous, conversational data in Dutch (Feng et al., 2022) and English (Adda-Decker and Lamel, 2005; Koenecke et al., 2020). They attribute these findings to sociolinguistic factors, like the higher incidence of disfluencies and informal speech in men.

Thus, the current literature provides a fragmented picture of gender disparities in speech technologies. However, most studies focus primarily on a single language, usually English, albeit across multiple datasets and models. We broaden this research to include large-scale *multi*lingual speech models. In addition, we evaluate performance in a third gender category, which includes individuals who do not identify as male or female.

3 Methodology

This section describes the experimental design to answer **Q1**, i.e., whether there are gendered performance differences in multi-task multilingual ASR models. Our results (§4) confirm it.

Models. We experiment with OpenAI's Whisper (Radford et al., 2022, WHISPER) and Meta's SeamlessM4T (Communication et al., 2023, SEAMLESS), two widely used, state-of-the-art multilingual ASR models (details in Appendix A.1).

Datasets. Among other multilingual datasets available for ASR (Gales et al., 2014; Black, 2019; Pratap et al., 2020; Iranzo-Sánchez et al., 2022; Valk and Alumäe, 2021), we use Mozilla Common Voice (`CV`, Ardila et al., 2020),[4] `Fleurs` (Conneau et al., 2023) and `VoxPopuli` (Wang et al., 2021). Mainly, dataset selection is bound to the availability of reliable speakers' gender information. We use the gender labels that come with each dataset.[5] Moreover, these datasets cover two distinct recording and speech conditions: *i) read*, where speech is typically well-articulated and based on the reading of pre-defined texts (`CV` and `Fleurs`), and *ii) spontaneous* conditions elicited from public speeches (`VoxPopuli`), which allow for more speaker-dependent articulations and variation in word usage (Gabler et al., 2023), and thus represent a testbed closer to real-world use cases of ASR technologies. We base our analysis on the concatenation of validation and test splits of each dataset to avoid unreliable results due to training data contamination. See Appendix A.5 for a discussion on data contamination and how transcription performance varies across splits in our setup.

Languages. We include 19 languages from `CV` and `Fleurs`, and a subset of those—11 in total, due to data availability—from `VoxPopuli`. They represent eight diverse language families and data availability conditions (i.e., high-low resource). The dataset choice depended on whether sufficient utterances stratified across genders were available for meaningful comparisons. See Tables 5-7, Appendix A.6 for all the statistics.

3.1 Evaluation

Quality Metrics. We use standard Word- and Character-Error Rate (WER and CER, respectively) to evaluate transcription quality. Following Radford et al. (2022), we *i*) report WER for all languages but Yoruba and Japanese—where variability in orthography and ambiguous word units may affect evaluation (Rowlands, 1954; Matsuoka et al., 1997)—for which we use CER, *ii*) transliterate all Russian and Serbian references and hypotheses into Cyrillic, and *iii*) apply the official normalization routines to texts when evaluating Whisper.[6] Moreover, we trim audio to the initial 30 seconds[7] and

[4] We use CV 16.0 from https://huggingface.co/datasets/mozilla-foundation/common_voice_16_0

[5] Speakers indicated gender identity with "Male" or "Female" labels in `Fleurs` and `CV`. The latter also includes an "Other" label. Gender is retrieved from https://multimedia.europarl.europa.eu/en in `VoxPopuli`.

[6] Found at https://github.com/openai/whisper/tree/main/whisper/normalizers

[7] Most recordings in our datasets are shorter than 30 seconds. See statistics in Table 5-7 in Appendix A.6.

filter out dataset noise by removing records with an empty reference and silence in the snippet.[8]

When evaluated in in terms of of overall quality, both WHISPER and SEAMLESS showcase competitive results, respectively, on `Fleurs` (12.68 and 19.87 – avg. 19 langs); `CV` (17.51 and 16.42 – avg. 19 langs); `VP` (13.59 and 12.74 – avg. 11 langs). For disaggregated error rate (ER) scores, see Figure 6 in Appendix A.1.

Gap Metrics. Measuring the ASR gender gap is equivalent to seeking what is commonly known in the field of machine learning as *group fairness* (Chouldechova and Roth, 2020), or, more specifically, *demographic parity* (Dwork et al., 2012). Conceptually, parity is achieved when a given statistical measure is equal across different groups.

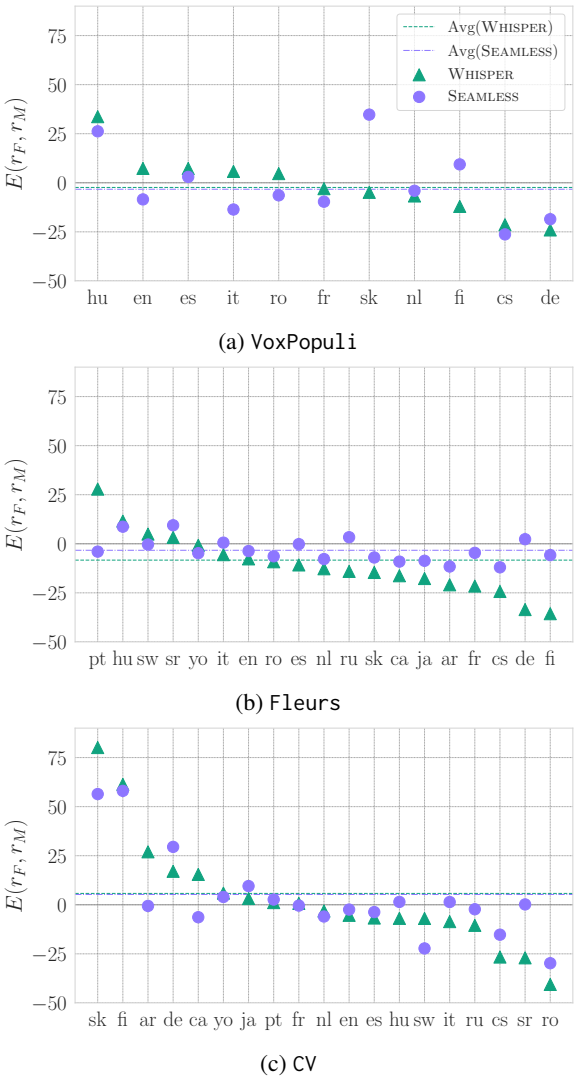
We operationalize demographic parity using the notation from Czarnowska et al. (2021). We introduce a Pairwise Comparison Metric, loosely inspired by the Disparity Score in Gaut et al. (2020), defined as follows:

$$E(r_A, r_B) = 100 \cdot (\phi(r_A) - \phi(r_B))/\phi(r_B) \quad (1)$$

where r_A and r_B are the set of audio snippets belonging to given groups (e.g., either "male", "female", or "other" in `CV`), and $\phi(\cdot)$ is one of the quality metrics defined above (i.e., WER or CER). The ideal score is 0—i.e., the model performs perfectly equally on the two groups.[9]

Significance test. We use a bootstrapped approach (Koehn, 2004; Søgaard et al., 2014) to estimate $\phi(\cdot)$. We sample 40% of the smallest group and the same number of records from the largest group for $n = 1000$ iterations. We sample stratifying on speakers to avoid skewing the speaker distribution. Stratified, gender-balanced sampling ensures a reliable comparison. We compute $E(\cdot)$ on the arithmetic means of $\phi(\cdot)$ across the n runs and use a two-sided Student's t-test to compute the statistical difference between the means. See Appendix A.5 for more details on comparing sampling performance vs. complete sets.

Note that sampling requires attributing each record to an individual speaker—information not available in `Fleurs`. For consistency, we attribute a speaker ID to each record automatically. Specifically, we *i*) use state-of-the-art speaker verification embedding models to encode each recording, *ii*) cluster them using HDBSCAN (Campello et al., 2013), and *iii*) assign a speaker ID to each cluster. See Appendix A.4 for details on the pipeline.



(a) `VoxPopuli`

(b) `Fleurs`

(c) `CV`

Figure 2: **Model error rate gap (Eq. 1).** Positive values indicate better performance on men, negative ones on women. $p < 0.05$.

4 Gender Performance Gaps

In the following, we report the results of evaluating WHISPER and SEAMLESS, focusing on gender performance disparities. We first investigate whether models are equally able to recognize female and male speakers (§4.1). Then, limited to `CV`, we repeat the analysis for speakers identifying with neither of the two against males (§4.2).

[8] When processing `Fleurs-es`, we noticed some records containing only background noise. We used state-of-the-art voice-activity detection models to detect and filter them out. See Appendix A.3 for details.

[9] ASR models achieve different baseline error rates across languages. Equation 1 measures relative, rather than absolute, gaps to account for this variability. See Table 3 in Appendix A.5 for details on absolute gaps and §7 for a discussion on possible metric interpretations.

4.1 "Female" – "Male" Gap

Figure 2 reports $E(r_F, r_M)$, i.e., the gender gap results for WHISPER and SEAMLESS on each dataset. Concerning our **Q1**, a broad overview reveals that these models do not perform equally across female and male speakers, often showing a preference for one gender over the other. This gender disparity in multilingual models is a key point of our investigation. Our analysis does not consistently reveal models disadvantaging the feminine group—see values below 0. This finding stands in stark contrast with mounting evidence of a strong masculine bias affecting a wide range of Natural Language Processing (NLP) tasks (e.g., Sun et al., 2019; Stanczak et al., 2022). It is especially noteworthy given the well-known under-representation of women in current resources used for model training (Garnerin et al., 2019; Zanon Boito et al., 2022; Sun and Peng, 2021). We observe that both speech models—albeit to varying degrees—better recognize female speech on average in two out of the three considered datasets (i.e., `Fleurs` and `VP`).

Upon closer examination, we also observe substantially different behaviors across datasets. First, despite the higher degree of spoken language variation to be expected in *spontaneous* recordings from the `VP` dataset (Figure 2a), results shows a comparatively reduced gender gap, with overlapping values between WHISPER and SEAMLESS for most languages (i.e., hu, es, fr, nl, cs, de). Conversely, it is on *read* resource that we attest to a higher degree of variability across models as well as languages, with occasional wide disparities toward either females or males in `CV` (2c). Notably, SEAMLESS remains closer to performance equality for most instances on `Fleurs` (2b), i.e., values between 11 and -10.

As such, in line with previous research (see §2), analysis of (binary) gender disparities for the ASR reveals a complex picture, which we further analyze and unpack in the following sections.

4.2 "Other" – "Male" Gap

Restricting our analysis to validation and test sets leaves us with few records for speakers who do not identify as either male or female. As such, we report results for five high-resource languages, Catalan, German, English, Spanish, and French.[10]

Table 1 reports $E(r_O, r_M)$ for both models. Results suggest that models penalize non-male speech and that gaps are larger than disparities observed for the female/male groups. Given the limited sample size, we deem this analysis only exploratory. However, it is a valuable first overview of the model's behaviors for gender-non-conforming voices. We underscore the need for more representation of diverse voices and sociodemographic groups in current resources, and for which monitoring fairness in existing speech models remains out of reach. See §7 for an expanded discussion.

	ca	de	en	es	fr
WHISPER	4.68	37.20	28.50	-2.00	5.21
SEAMLESS	52.76	17.19	38.70	11.37	-0.50

Table 1: **Model error rate gap (Eq. 1).** Positive values indicate better performance on men, negative ones on "Other." $p < 0.05$.

5 Acoustic and Lexical Analysis

Gendered performance gaps vary across languages and datasets. Here, we dive into a focused analysis on acoustic aspects or lexical phenomena present in the test data (**Q2**). Due to data availability, we conduct this analysis on the "Female" - "Male" setup.

5.1 Acoustic Analysis

Motivated by sociophonetic evidence for gendered differences in speech (§2), we examine voice and language production across speakers of different genders. Specifically, we measure three acoustic features in our evaluation records, and explore whether potential acoustic differences across gender groups relate to gender performance gaps.

We include: *i*) *pitch*, measured as the mean of the fundamental frequency values (Hirst and de Looze, 2021), known to vary between biological sexes (Coleman, 1976; Hillenbrand and Clark, 2009) and languages (Loveday, 1981; Yuasa, 2008); *ii*) *intensity*, measured as the mean of fundamental frequencies (Pausewang Gelfer and Young, 1997), subject to recording conditions (Maryn and Zarowski, 2015); *iii*) *speaking rate*, measured as the number of tokens per minute (Künzel, 2013), subject to language and gender variation (Borsel and Maesschalck, 2008; Coupé et al., 2019) as well as read vs. spontaneous speech (Nakamura et al., 2008).[11]

[10] Size varies from 8 (`ca`, min) to 86 (`es`, max) records.

[11] We compute average pitch and intensity using Praat (Boersma and Weenink, 2001). Concerning speaking rate, we found no reliable multilingual tools for syllable segmentation. Therefore, we approximated it by the number of tokens as computed by WHISPER's tokenizer.

Findings. We started by exploring whether pitch, intensity, and speaking rate differ between genders in each dataset-language pair. Independent-sample T-tests revealed significant differences for most setups ($p < 0.05$), consistent with previous literature. For example, pitch consistently showed statistically significant differences (see Appendix B.1).

Motivated by these observations, we computed the mean acoustic values of each gender group and correlated it with performance gaps, i.e., $E(r_F, r_M)$, aggregating by dataset and model. No clear trends emerged overall. Sporadically, we found a strong and significant linear correlation, e.g., for SEAMLESS and VoxPopuli where Pearson's *rho* for pitch is -0.83 (see Appendix B.1 for complete details). Fitting an ordinary least squares (OLS) regressor to predict $E(r_F, r_M)$ led to similar conclusions, as low R^2 scores suggest (max: 0.42, average$_\sigma$: $0.24_{\pm 0.11}$). For a more fine-grained analysis, we repeated the analysis fitting an OLS model to predict sentence-level error rates (i.e., r_F, r_M) but found low R^2 scores (max: 0.20, average$_\sigma$: $0.03_{\pm 0.04}$).

In summary, despite pitch, intensity, and speaking rate vary significantly across gender groups within the same dataset and language, such variation seldom correlates with performance differences. These findings suggest that performance gaps are a complex phenomenon, and our analysis should expand elsewhere to explain it.

5.2 Lexical Analysis

In §5.1, we explored acoustic aspects of speech. However, *what* is uttered can be as crucial as *how* it is uttered. Indeed, prior work has found that overall speech perception can be impacted by aspects related to lexical and syntactic complexities (van Knijff et al., 2018; Carroll and Ruigendijk, 2012). Besides, certain lexical phenomena such as named entities represent a well-known challenge for speech models, especially in multilingual contexts (Gaido et al., 2021). Thus, we focus on speech content and study whether lexical phenomena—as measured in the reference transcripts—explain ASR disparities.

For each record, we counted *i*) the occurrences of part-of-speech tags and *ii*) named entities, and computed *iii*) lexical density (Halliday, 1989). Similarly to acoustic features, we contrast distribution between gender groups (details in Appendix B.2).

Findings. Comparing the distribution of lexical features between female and male speakers, we found mixed results. Several dataset-language setups have significantly different distributions ($p < 0.05$), but not all. Different distributions are also present in read datasets, suggesting that lexical variability is not controlled when collecting data. We proceeded similarly to the acoustic analysis to verify whether such differences explain error rates. Fitting an OLS model to predict error rates from lexical features only yields low R^2 scores (max: 0.35, average$_\sigma$: $0.09_{\pm 0.08}$, across all datasets and languages). Moreover, we found no significant (linear) correlation between the difference of group means and $E(r_F, r_M)$.

This finding echoes those from acoustic analysis: Lexical phenomena do not explain gender performance gaps in our data. We must research other aspects beyond the lexical content of utterances.

6 Probing Gender in ASR Models

The field of natural language processing (NLP) has now established that transformer language models encode syntactic (Hewitt and Manning, 2019), semantic (Tenney et al., 2019), and factual (Petroni et al., 2019; Meng et al., 2022) information in hidden representations. As such, recent work has focused on *extracting* this information through probes, i.e., supervised classifiers trained on the model's embeddings (Alain and Bengio, 2016; Belinkov and Glass, 2019). Compellingly, some have related extractability of sensitive attributes, e.g., gender, to bias in downstream application (Orgad et al., 2022). Probing is primarily motivated by the risk that models entangle protected attributes and predictions in sensitive use cases (Zhao et al., 2018; Ravfogel et al., 2020).

Motivated by this line of research, we measure gender extractability in ASR models and ask (**Q2**) whether and to what extent it explains gender-based performance gaps. To our knowledge, ours is the first study of this kind.[12]

Experimental design. We focus on one dataset-model configuration, namely WHISPER and CV. We attach our probes to the model encoder's last layer embeddings and train one distinct probe for every position. We use Logistic Regression and Mini-

[12] Interpretability for speech models is a relatively new research avenue. Mohebbi et al. (2023) use context-mixing techniques to explain homophone disambiguation; Pastor et al. (2023) mask word-units to explain intent detection models. Following our work, Krishnan et al. (2024) use amnesic probing to linearly erase gender information in transformer-based ASR models, examining its effects on downstream performance.

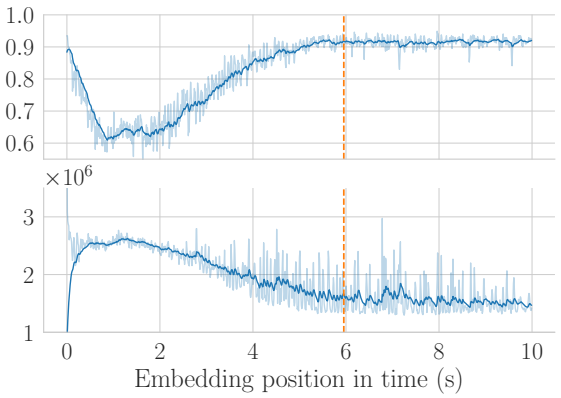


Figure 3: **F1 Macro (top) and measured code length (bottom)** for Logistic Regression and MDL probes, respectively. English, F-M setup, first 10 seconds of context. Actual score (blue pale line), exponential moving average (solid line, n=3), and average length of test snippets (dashed line).

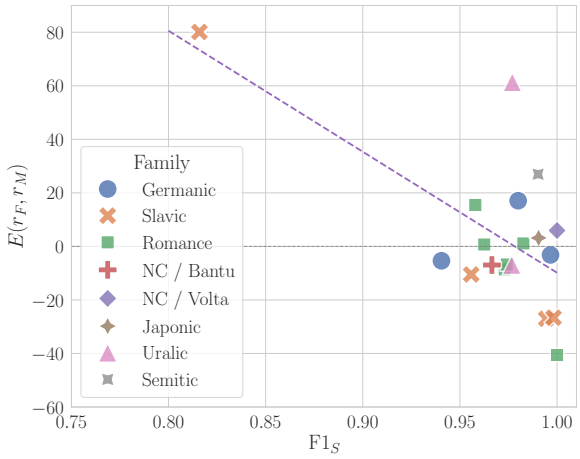


Figure 4: **Logistic probe $F1_S$ vs.** $E(r_F, r_M)$. CV. The purple dashed line is a linear interpolation.

mum Description Length (MDL) (Voita and Titov, 2020) to probe gender in the female-male binary setup (see Appendix B.3 for details).

Findings. Figure 3 shows gender extractability on the English CV split for the F-M setup. Trends are similar across languages (full results in Figure 9a). First, F1 scores suggest that **gender extractability is relatively easy**. Whisper produces representations of female and male speech that can be easily separated. Second, extractability is not constant over time. It starts high in the first milliseconds, drops during the actual signal (i.e., the speaker is talking), and finally plateaus around the initial value. This finding recalls the "attention sink" theory, implicating that transformers prioritize initial positions for specific tasks (Xiao et al., 2023). Third, the trends observed in MDL probes—where lower codelength indicates an easier task and higher extractability—suggest that simple logistic regression is a robust method, countering the issues identified in NLP (Voita and Titov, 2020).[13]

Correlation with Error Rate. Do gender probing scores tell us something about ASR quality and gender bias (**Q2**)? To answer this question, we compare F1 scores from logistic regression scores and ASR error rates. We measure $F1_S$, i.e., the average F1 score achieved by probes at the positions corresponding to actual speech,[14] $\phi(r_M)$ and $\phi(r_F)$ (error rates on male and female segments), and the disparity score $E(r_F, r_M)$.

$F1_S$ has a weak linear correlation with $\phi(r_M)$ and $\phi(r_F)$ (Pearson's ρ is 0.13 and -0.003, respectively), hinting that $F1_S$ cannot explain per-group quality. However, ρ between $F1_S$ and $E(r_F, r_M)$ is -0.65 ($n = 19$, see Figure 4). In other words, linear correlation suggests that **the better probes can extract gender from WHISPER's hidden states, the lower the F-M error rate gap is**. Note that, as per its definition, $E(r_F, r_F)$ can be negative (see points below the zero line in Figure 4). Negative values indicate performance favoring the minority group (here, women). This finding suggests that WHISPER *does* encode recordings from speakers that identify as men and women differently. This aspect can serve as a proxy for measuring and mitigating gender disparities, e.g. by reducing gender extractability from hidden representations (Krishnan et al., 2024). However, we caution the reader from attributing high extractability to gendered voice and discourage using probes on people's voice to predict gender (see Ethical considerations, §8).

7 Discussion

Twists, humps, and pebbles—the way to equitable multilingual speech recognition models is no straight line. We have discovered that gender disparities vary across spoken setups and languages and that consistency is also weak between models. Perhaps of greater interest, the advantaged group is sometimes women, sometimes men, and never

[13] To confirm this hypothesis, we conduct additional experiments by training logistic probes with randomly shifted labels. We report random guess performance on all languages (see Figure 9b) in Appendix B.3.

[14] We approximate this interval by considering all positions before the average test set length (orange line in Figure 3).

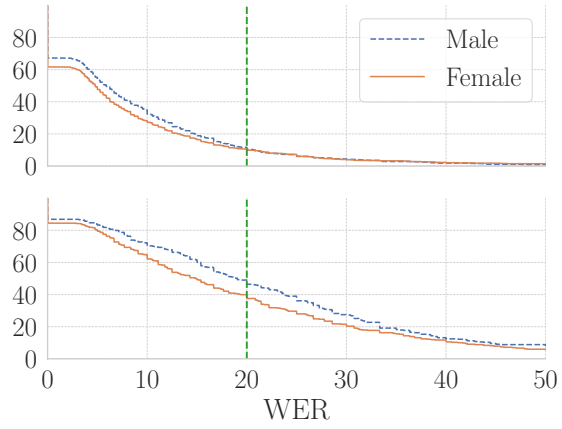


Figure 5: **Share of records (y axis, %) having a WER greater than the value on the x axis.** WHISPER, `Fleurs` nl (top) and ar (bottom).

anyone who does not identify with either.

Such a high variability prevents us from foreseeing gender disparities and, in turn, the actual impact on system users. We discuss here how to *i*) reduce such variability and *ii*) better use the metrics and test data at our disposal.

Free! the training data. While our acoustic and lexical inquiries on test data were inconsequential, experiments with probes on the model's internals showed that models learned and encoded properties to differentiate genders. Intuitively, a follow-up study would focus on what shaped those model's properties in the first place: training data. Models encode facts and information from training data, and some studies looked into how different sociodemographic groups are represented (Muller et al., 2023; Elazar et al., 2024). Studying gender distributions in pretraining multilingual speech data would be extremely valuable, as we could establish how gendered gaps and underrepresentation correlate. However, this analysis was impossible, as neither Whisper (Radford et al., 2022) nor SeamlessM4T (Communication et al., 2023) released such information. We call for more transparency.

Go beyond group-wise metrics. A crucial choice in our design was measuring differences between two groups of speakers. We opted for $E(\cdot)$, measured as the *relative* difference between the group means. While the metric was primarily meant to make results across datasets, languages, and models comparable, it does not provide a complete picture of gender disparity's concrete impact. For instance, it hides *absolute* gaps: a $E(\cdot) = 10$ gap corresponds to a 0.2 WER gap if the model averages a WER of 2 on male speakers and to a 2 WER gap if the average is 20. These gaps may or may not be significant given many contextual factors (e.g., what the model is being used for, who is using it, and which ramifications such gaps can have). This thinking echoes that of recent critiques on measuring social bias in NLP, advocating for explicit statements about what system behavior must be considered good or bad and under which social values (Blodgett et al., 2020), and how to build contextualized metrics (Lum et al., 2024).

We would like a metric to indicate a more precise estimation of ASR model failures and their severity for users. A step in this direction would be going beyond group-wise metrics and studying sentence-level errors. Drawing on the concepts of wealth distribution and inequality (O'sullivan et al., 2003), we can inspect error distributions between speakers and groups and address the question "*assuming X is an acceptable error rate, what fraction of records from each group will be served to a satisfactory level?*" Following Koenecke et al. (2020), we report in Figure 5 the proportion of records having an error rate at least that large. Due to space constraints, we focus on WHISPER's transcriptions of `Fleurs`' Arabic and Dutch data. This visualization lets us identify where performance between genders starts to diverge. If, given our use case, we hypothesize that transcriptions become unusable for WER > 20, men and women would be equally affected in Dutch. However, it would not be the case in Arabic where 40% of records from female speakers, but half of those from males, would not be acceptable. Parity is reached around WER=40.

This concrete analysis highlighted gender disparities that group-wise metrics could not capture otherwise. It also served as an example of a broader argumentation: Future inquiries can benefit from using metrics and observations more closely related to real-world scenarios.

Improve sociodemographic representation. Empirical studies like ours rely on the quality and quantity of evaluation data. To test speech recognition, we need representative splits from diverse speakers. However, current datasets are gender-skewed, with men often overrepresented in terms of records and speakers. While downsampling can help at evaluation time, it does not address the lack of speaker diversity. Skewed distributions are even more evident in

low-resource languages when building standard train/validation/test sets is required—e.g., in CV, Yoruba counts only 14 and 20 unique males and females, respectively. Catalan counts only 8 within "Other."

Therefore, we call for much needed considerations when comparing gender groups in high-resource languages, and new collective efforts to collect more representative data for low-resource languages and gender identities beyond the binary.

8 Conclusion

We conducted the first extensive evaluation on gender-based performance gaps of Whisper (Radford et al., 2022) and SeamlessM4T (Communication et al., 2023) for ASR. These models consistently exhibit gender bias across 19 languages from eight language families. Depending on the language, disparities can favor men or women but rarely favor speakers who identify as neither. We locate a potential source of these gaps using gender probes from interpretability approaches. Our results show that probes can be a proxy for gender gaps and that group fairness in multi-task multilingual models remains unsolved.

Ethical Considerations

The use of gender as a variable in this paper warrants ethical reflections.

As one of the most salient perceptual traits of one's identity (Kreiman and Sidtis, 2011; Azul, 2015; Zimman, 2021) gendered differences represent a linchpin of much (socio)phonetic research (Zimman, 2020), which has unpacked several physical and sociocultural factors contributing to such gendered characteristics in the voice. Based on this evidence, our work does not intend to be normative nor assumes the existence of a single, unidimensional "female" or "male" voice. Instead, we question whether different gendered groups are equally recognized by current multilingual ASR models and incorporate sociophonetic knowledge in our analysis and discussions to isolate why that might not be the case. To do so, we do not make any inference about the gender of the speakers in the employed data. Instead, we rely on the declared gender of the speakers in the employed speech resources. In this regard, part of our analysis of the Mozilla Common Voice dataset uses a third category, "Other", which potentially aggregates diverse identities (e.g., transgender, non-binary, and other marginalized individuals)[15] under one single umbrella term. While this third category also allows us to include genders non-conforming to the binary in our study, we recognize that this label and category might be an oversimplification and risk erasing the experiences and representativity of many gender identities.

Finally, gender probes represent a methodological approach to studying how models encode different audios from input signals. It is, hence, essential to remember the inherent limitations and ethical implications of relying on probes for gender classification. While these tools may offer a convenient means of categorization and can be suitable for exploring the models' behavior, they often overlook the nuanced and multifaceted nature of gender identity.

Limitations

Our paper comes with a series of limitations. We divide them into two categories: data and methods.

Data. Gender is the driving variable of our analysis. However, it is widely recognized that gender interacts with other sociocultural factors, e.g., dialect (Wolfram, 2004) or sexual orientation (Zimman, 2013), in voice production. By limiting to self-identified gender provided with the datasets, our analysis can only provide a partial view. We thus advocate for speech dataset releases that include a principled set of such factors.

The scarcity of data limits the generalizability of our results in two more setups in our analysis. First, the category "Other" counts a deficient number of speakers and records in most of our setups (see §7). Second, the number of speakers we extracted from Fleurs is compared to CV and VoxPopuli (see 3.1). These factors hamper cross-dataset comparison and overall generalizability.

We did not control for data quality on references. It might be that references are noisy (e.g., single words, empty) and can lead to over- or underestimation of our measurements.

Finally, data contamination. Although we chose validation and test sets as our analysis targets, we cannot exclude that models were trained on part or all of them. We conducted a side analysis (Appendix A.5) that suggested that this *might* not be the case.

[15] These are among the offered gender options that can currently be reported when donating one's voice for Common Voice.

Methods. Since we miss reliable multilingual tools for acoustic analyses, we estimated the speaking rate using Whisper's pre-trained tokenizer. We acknowledge that pretrained tokenizers yield fewer tokens for high-resource languages. Hence, our measurement of low-resource might be overestimating the phenomenon.

If framed in the context of bias evaluation paradigms—which we are currently not doing—probing can be seen as an intrinsic bias paradigm. We correlate probing performance to a downstream task (and potential harm), i.e., quality gaps in ASR. However, studies in the field of NLP recognize that intrinsic and extrinsic (downstream) bias metrics do not necessarily correlate (Goldfarb-Tarrant et al., 2021; Kaneko et al., 2022). To our knowledge, ours is the first study to inspect such aspects in transformers for speech, and findings may not hold across modalities. We leave this research question to future work.

Acknowledgments

We thank the reviewers, the members of the SARDINE and MilaNLP research groups, and Marco Gaido for the insightful comments. Giuseppe Attanasio was supported by the Portuguese Recovery and Resilience Plan through project C645008882-00000055 (Center for Responsible AI) and by Fundação para a Ciência e Tecnologia through contract UIDB/50008/2020. He conducted part of the work as a member of the MilaNLP group at Bocconi University, Milan. Dirk Hovy was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 949944, INTEGRATOR) and a MUR FARE 2020 initiative under grant agreement Prot. R20YSMBZ8S (INDOMITA). He is director of the Data and Marketing Insights Unit of the Bocconi Institute for Data Science and Analysis (BIDSA). Beatrice Savoldi was supported by the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU.

References

Martine Adda-Decker and Lori Lamel. 2005. Do speech recognizers prefer female speakers? In *Proc. Interspeech 2005*, pages 2205–2208.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

H Samy Alim. 2004. *You know my steez: An ethnographic and sociolinguistic study of styleshifting in a Black American speech community*. Stanford University.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222.

David Azul. 2015. On the varied and complex factors affecting gender diverse people's vocal situations: Implications for clinical practice. *Perspectives on Voice and Voice Disorders*, 25(2):75–86.

Kara Becker, Sameer ud Dowla Khan, and Lal Zimman. 2022. Beyond binary gender: creaky voice, gender, and the variationist enterprise. *Language Variation and Change*, 34(2):215–238.

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72.

Michael Beukman and Manuel Fokam. 2023. Analysing cross-lingual transfer in low-resourced African named entity recognition. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 199–224, Nusa Dua, Bali. Association for Computational Linguistics.

Alan W Black. 2019. Cmu wilderness multilingual speech dataset. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Paul Boersma and David Weenink. 2001. Praat, a system for doing phonetics by computer. *Glot International*, 5(9/10):341–345.

John Van Borsel and Dorothy De Maesschalck. 2008. Speech rate in males, females, and male-to-female transsexuals. *Clinical Linguistics & Phonetics*, 22(9):679–685. PMID: 18608249.

Peter A Busby and Geoff L Plant. 1995. Formant frequency values of vowels produced by preadolescent boys and girls. *The Journal of the Acoustical Society of America*, 97(4):2603–2606.

Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172. Springer.

Rebecca Carroll and Esther Ruigendijk. 2012. The effects of syntactic complexity on processing sentences in noise. *Journal of Psycholinguistic Research*, 42:139 – 159.

Alexandra Chouldechova and Aaron Roth. 2020. A snapshot of the frontiers of fairness in machine learning. *Communications of the ACM*, 63(5):82–89.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Ralph O. Coleman. 1976. A comparison of the contributions of two voice quality characteristics to the perception of maleness and femaleness in the voice. *Journal of Speech & Hearing Research*, 19(1):168–180.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, John Hoffman, Min-Jae Hwang, Hirofumi Inaguma, Christopher Klaiber, Ilia Kulikov, Pengwei Li, Daniel Licht, Jean Maillard, Ruslan Mavlyutov, Alice Rakotoarison, Kaushik Ram Sadagopan, Abinesh Ramakrishnan, Tuan Tran, Guillaume Wenzek, Yilin Yang, Ethan Ye, Ivan Evtimov, Pierre Fernandez, Cynthia Gao, Prangthip Hansanti, Elahe Kalbassi, Amanda Kallet, Artyom Kozhevnikov, Gabriel Mejia Gonzalez, Robin San Roman, Christophe Touret, Corinne Wong, Carleigh Wood, Bokai Yu, Pierre Andrews, Can Balioglu, Peng-Jen Chen, Marta R. Costa-jussà, Maha Elbayad, Hongyu Gong, Francisco Guzmán, Kevin Heffernan, Somya Jain, Justine Kao, Ann Lee, Xutai Ma, Alex Mourachko, Benjamin Peloquin, Juan Pino, Sravya Popuri, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, Anna Sun, Paden Tomasello, Changhan Wang, Jeff Wang, Skyler Wang, and Mary Williamson. 2023. Seamless: Multilingual Expressive and Streaming Speech Translation.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *2022 IEEE Spoken Language Technology Workshop (SLT)*, pages 798–805. IEEE.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):eaaw2594.

Paula Czarnowska, Yogarshi Vyas, and Kashif Shah. 2021. Quantifying social biases in NLP: A generalization and empirical comparison of extrinsic fairness metrics. *Transactions of the Association for Computational Linguistics*, 9:1249–1267.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226.

Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's in my big data? In *The Twelfth International Conference on Learning Representations*.

Rui Feng, Chen Luo, Qingyu Yin, Bing Yin, Tuo Zhao, and Chao Zhang. 2022. CERES: Pretraining of graph-conditioned transformer for semi-structured session data. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 219–230, Seattle, United States. Association for Computational Linguistics.

Dennis Fucci, Marco Gaido, Matteo Negri, Mauro Cettolo, and Luisa Bentivogli. 2023. No pitch left behind: Addressing gender unbalance in automatic speech recognition through pitch manipulation. *arXiv preprint arXiv:2310.06590*.

Susanne Fuchs and Martine Toda. 2010. Do differences in male versus female/s/reflect biological or sociophonetic factors. *Turbulent sounds: An interdisciplinary guide*, 21:281–302.

Philipp Gabler, Bernhard C Geiger, Barbara Schuppler, and Roman Kern. 2023. Reconsidering Read and Spontaneous Speech: Causal Perspectives on the Generation of Training Data for Automatic Speech Recognition. *Information*, 14(2):137.

Marco Gaido, Susana Rodríguez, Matteo Negri, Luisa Bentivogli, and Marco Turchi. 2021. Is "moby dick" a whale or a bird? named entities and terminology in speech translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1707–1716.

Marco Gaido, Beatrice Savoldi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2020. Breeding gender-aware direct speech translation systems. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3951–3964, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Mark JF Gales, Kate M Knill, Anton Ragni, and Shakti P Rath. 2014. Speech recognition and keyword spotting for low-resource languages: Babel project research at cued. In *Fourth International workshop on spoken language technologies for under-resourced languages (SLTU-2014)*, pages 16–23. International Speech Communication Association (ISCA).

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2019. Gender Representation in French Broadcast Corpora and Its Impact on ASR Performance. In *1st International Workshop on AI for Smart TV Content Production, Access and Delivery*, page 3–9.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2020. Gender representation in open source speech resources. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6599–6605, Marseille, France. European Language Resources Association.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2021. Investigating the Impact of Gender Representation in ASR Training Data: a Case Study on Librispeech. In *3rd Workshop on Gender Bias in Natural Language Processing*, pages 86–92.

Andrew Gaut, Tony Sun, Shirlyn Tang, Yuxin Huang, Jing Qian, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2020. Towards understanding gender bias in relation extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2943–2953, Online. Association for Computational Linguistics.

Seraphina Goldfarb-Tarrant, Rebecca Marchant, Ricardo Muñoz Sánchez, Mugdha Pandya, and Adam Lopez. 2021. Intrinsic bias metrics do not correlate with application bias. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1926–1940, Online. Association for Computational Linguistics.

Michael A.K. Halliday. 1989. Spoken and written language.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

James M. Hillenbrand and Michael J. Clark. 2009. The role of f0 and formant frequencies in distinguishing the voices of men and women. *Attention Perception & Psychophysics*, 71(5):1150–1166.

Daniel J. Hirst and Céline de Looze. 2021. Measuring Speech. Fundamental frequency and pitch. In Rachael-Anne Knight and Jane Setter, editors, *Cambridge Handbook of Phonetics*, 1, pages 336–361. Cambridge University Press.

Aliakbar Imani and Hadina Habil. 2017. Lexical features of academic writing.

Javier Iranzo-Sánchez, Javier Jorge Cano, Alejandro Pérez-González-de Martos, Adrián Giménez Pastor, Gonçal Garcés Díaz-Munío, Pau Baquero-Arnal, Joan Albert Silvestre-Cerdà, Jorge Civera Saiz, Albert Sanchis, and Alfons Juan. 2022. MLLP-VRAIN UPV systems for the IWSLT 2022 simultaneous speech translation and speech-to-speech translation tasks. In *Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022)*, pages 255–264, Dublin, Ireland (in-person and online). Association for Computational Linguistics.

Masahiro Kaneko, Danushka Bollegala, and Naoaki Okazaki. 2022. Debiasing isn't enough! – on the effectiveness of debiasing MLMs and their social biases in downstream tasks. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 1299–1310, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Philipp Koehn. 2004. Statistical significance tests for machine translation evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 388–395, Barcelona, Spain. Association for Computational Linguistics.

Allison Koenecke, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. 2020. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences*, 117(14):7684–7689.

Jody Kreiman and Diana Sidtis. 2011. *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

Aravind Krishnan, Badr M. Abdullah, and Dietrich Klakow. 2024. On the encoding of gender in transformer-based asr representations.

Hermann J. Künzel. 2013. Some general phonetic and forensic aspects of speaking tempo. *International Journal of Speech Language and The Law*, 4:48–83.

William Labov. 1964. *The social stratification of English in New York city*. Ph.D. thesis, Columbia University.

Fangfang Li. 2017. The development of gender-specific patterns in the production of voiceless sibilant fricatives in Mandarin Chinese. *Linguistics*, 55(5):1021–1044.

Chunxi Liu, Michael Picheny, Leda Sarı, Pooja Chitkara, Alex Xiao, Xiaohui Zhang, Mark Chou, Andres Alvarado, Caner Hazirbas, and Yatharth Saraf. 2022.

Towards measuring fairness in speech recognition: Casual conversations dataset transcriptions. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6162–6166.

Leo Loveday. 1981. Pitch, politeness and sexual role: An exploratory investigation into the pitch correlates of English and Japanese politeness formulae. *Language and Speech*, 24(1):71–89.

Kristian Lum, Jacy Reese Anthis, Chirag Nagpal, and Alexander D'Amour. 2024. Bias in language models: Beyond trick tests and toward ruted evaluation. *arXiv preprint arXiv:2402.12649*.

Youri Maryn and Andrzej Zarowski. 2015. Calibration of clinical audio recording and analysis systems for sound intensity measurement. *American Journal of Speech-Language Pathology*, 24(4):608–618.

Tatsuo Matsuoka, Katsutoshi Ohtsuki, Takeshi Mori, Kotaro Yoshida, Sadaoki Furui, and Katsuhiko Shirai. 1997. Japanese large-vocabulary continuous-speech recognition using a business-newspaper corpus. In *1997 IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 3, pages 1803–1806. IEEE.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in Neural Information Processing Systems*, 35:17359–17372.

Zion Mengesha, Courtney Heldreth, Michal Lahav, Juliana Sublewski, and Elyse Tuennerman. 2021. "I don't Think These Devices are Very Culturally Sensitive."—Impact of Automated Speech Recognition Errors on African Americans. *Frontiers in Artificial Intelligence*, 4.

Josh Meyer, Lindy Rauchenstein, Joshua D Eisenberg, and Nicholas Howell. 2020. Artie bias corpus: An open dataset for detecting demographic bias in speech applications. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6462–6468.

Hosein Mohebbi, Grzegorz Chrupała, Willem Zuidema, and Afra Alishahi. 2023. Homophone disambiguation reveals patterns of context mixing in speech transformers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8249–8260, Singapore. Association for Computational Linguistics.

Benjamin Muller, Belen Alastruey, Prangthip Hansanti, Elahe Kalbassi, Christophe Ropers, Eric Smith, Adina Williams, Luke Zettlemoyer, Pierre Andrews, and Marta R. Costa-jussà. 2023. The gender-GAP pipeline: A gender-aware polyglot pipeline for gender characterisation in 55 languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 536–550, Singapore. Association for Computational Linguistics.

Masanobu Nakamura, Koji Iwano, and Sadaoki Furui. 2008. Differences between acoustic characteristics of spontaneous and read speech and their effects on speech recognition performance. *Computer Speech & Language*, 22(2):171–184.

Jennifer Oates and Georgia Dacakis. 2015. Transgender voice and communication: Research evidence underpinning voice intervention for male-to-female transsexual women. *Perspectives on Voice and Voice Disorders*, 25(2):48–58.

Hadas Orgad, Seraphina Goldfarb-Tarrant, and Yonatan Belinkov. 2022. How gender debiasing affects internal model representations, and why it matters. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2602–2628, Seattle, United States. Association for Computational Linguistics.

Arthur O'sullivan, Steven M Sheffrin, and Kathy Swan. 2003. Economics: Principles in action.

Eliana Pastor, Alkis Koudounas, Giuseppe Attanasio, Dirk Hovy, and Elena Baralis. 2023. Explaining speech classification models via word-level audio segments and paralinguistic features. *arXiv preprint arXiv:2309.07733*.

Marylou Pausewang Gelfer and Shannon Ryan Young. 1997. Comparisons of intensity measures and their stability in male and female sneakers. *Journal of Voice*, 11(2):178–186.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Nicolai Pharao, Marie Maegaard, Janus Spindler Møller, and Tore Kristiansen. 2014. Indexical meanings of [s] among Copenhagen youth: Social perception of a phonetic variant in different prosodic contexts. *Language in Society*, 43(1):1–31.

Robert J. Podesva, Janneke Van Hofwegen, Erez Levon, and Ronald Beline Mendes. 2016. S/exuality in small-town California: Gender normativity and the acoustic realization of/s. *Language, sexuality, and power: Studies in intersectional linguistics*, pages 16–88.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust Speech Recognition via Large-Scale Weak Supervision. ArXiv:2212.04356 [cs, eess].

Sai Sathiesh Rajan, Sakshi Udeshi, and Sudipta Chattopadhyay. 2022. Aequevox: Automated fairness testing of speech recognition systems. In *International Conference on Fundamental Approaches to Software Engineering*, pages 245–267. Springer International Publishing Cham.

Mirco Ravanelli, Titouan Parcollet, Peter Plantinga, Aku Rouhe, Samuele Cornell, Loren Lugosch, Cem Subakan, Nauman Dawalatabad, Abdelwahab Heba, Jianyuan Zhong, Ju-Chieh Chou, Sung-Lin Yeh, Szu-Wei Fu, Chien-Feng Liao, Elena Rastorgueva, François Grondin, William Aris, Hwidong Na, Yan Gao, Renato De Mori, and Yoshua Bengio. 2021. SpeechBrain: A general-purpose speech toolkit. ArXiv:2106.04624.

Shauli Ravfogel, Yanai Elazar, Hila Gonen, Michael Twiton, and Yoav Goldberg. 2020. Null it out: Guarding protected attributes by iterative nullspace projection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7237–7256, Online. Association for Computational Linguistics.

Evan C Rowlands. 1954. Types of word junction in yoruba. *Bulletin of the School of Oriental and African Studies*, 16(2):376–388.

Majdi Sawalha and Mohammad Abu Shariah. 2013. The effects of speakers' gender, age, and region on overall performance of Arabic automatic speech recognition systems using the phonetically rich and balanced Modern Standard Arabic speech corpus. In *Proceedings of the 2nd Workshop of Arabic Corpus Linguistics WACL-2*. Leeds.

Adrian P. Simpson. 2009. Phonetic differences between male and female speech. *Language and linguistics compass*, 3(2):621–640.

Anders Søgaard, Anders Johannsen, Barbara Plank, Dirk Hovy, and Hector Martínez Alonso. 2014. What's in a p-value in NLP? In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 1–10, Ann Arbor, Michigan. Association for Computational Linguistics.

Karolina Stanczak, Edoardo Ponti, Lucas Torroba Hennigen, Ryan Cotterell, and Isabelle Augenstein. 2022. Same neurons, different languages: Probing morphosyntax in multilingual pre-trained models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1589–1598, Seattle, United States. Association for Computational Linguistics.

Jiao Sun and Nanyun Peng. 2021. Men are elected, women are married: Events gender bias on wikipedia. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 350–360.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating gender bias in natural language processing: Literature review. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1630–1640, Florence, Italy. Association for Computational Linguistics.

Tina Tallon. 2019. A Century of "Shrill": How Bias in Technology Has Hurt Women's Voices. *The New Yorker*.

Rachael Tatman. 2017. Gender and dialect bias in YouTube's automatic captions. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 53–59, Valencia, Spain. Association for Computational Linguistics.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. BERT rediscovers the classical NLP pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Jörgen Valk and Tanel Alumäe. 2021. Voxlingua107: a dataset for spoken language recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 652–658. IEEE.

Eline C van Knijff, Martine Coene, and Paul J. Govaerts. 2018. Speech understanding in noise in elderly adults: the effect of inhibitory control and syntactic complexity. *International journal of language & communication disorders*, 53 3:628–642.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196, Online. Association for Computational Linguistics.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003, Online. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Walt Wolfram. 2004. Urban african american vernacular english: Morphology and syntax. *A handbook of varieties of English*, 1:319–340.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Ikuko Patricia Yuasa. 2008. *Culture and gender of voice pitch: A sociophonetic comparison of the Japanese and Americans*. Equinox Publishing.

Marcely Zanon Boito, Fethi Bougares, Florentin Barbier, Souhir Gahbiche, Loïc Barrault, Mickael Rouvier, and Yannick Estève. 2022. Speech resources in the Tamasheq language. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2066–2071, Marseille, France. European Language Resources Association.

Jieyu Zhao, Yichao Zhou, Zeyu Li, Wei Wang, and Kai-Wei Chang. 2018. Learning gender-neutral word embeddings. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4847–4853, Brussels, Belgium. Association for Computational Linguistics.

Lal Zimman. 2013. Hegemonic masculinity and the variability of gay-sounding speech: The perceived sexuality of transgender men. *Journal of Language and Sexuality*, 2(1):1–39.

Lal Zimman. 2017. Gender as stylistic bricolage: Transmasculine voices and the relationship between fundamental frequency and/s. *Language in Society*, 46(3):339–370.

Lal Zimman. 2018. Transgender voices: Insights on identity, embodiment, and the gender of the voice. *Language and Linguistics Compass*, 12(8):e12284.

Lal Zimman. 2020. Sociophonetics. *The International Encyclopedia of Linguistic Anthropology*, pages 1–5.

Lal Zimman. 2021. Gender diversity and the voice. In *The Routledge handbook of language, gender, and sexuality*, pages 69–90. Routledge.

A Experimental Details

A.1 Multilingual ASR Models

For both Whisper and SeamlessM4T, we used code and model checkpoints in `transformers` (Wolf et al., 2020). The Hub's model IDs are openai/whisper-large-v3 and facebook/seamless-m4t-v2-large, respectively. Both checkpoints correspond to the latest and best-performing versions available at the time of writing, February 2024.

We used each models's standard decoding configurations to transcribe each audio snippet.

In figure 6, we report WER results for SEAMLESS and WHISPER on the languages and datasets used in our experiments.

A.2 Languages

We studied 19 languages in `CV` and `Fleurs` and 11 in `VoxPopuli`. Languages cover eight distinct language families. Due to data availability, we limited the comparison between speaker identifying with "Male" and "Other" in `CV` to Catalan, German, English, Spanish, and French. Table 4 reports an overview of languages and language families.

A.3 Voice Activity Detection

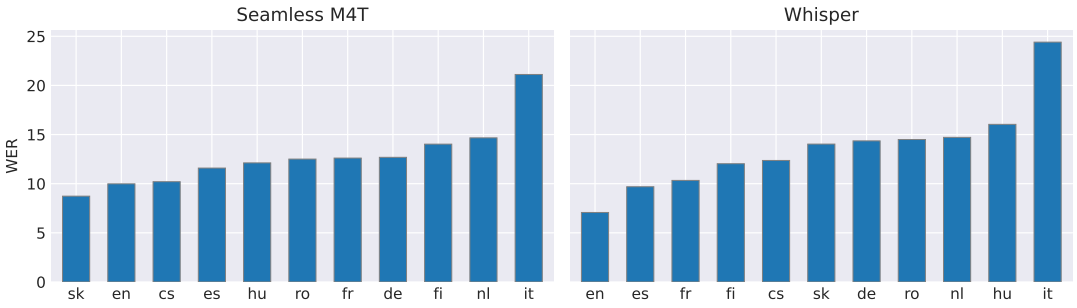
While conducting our experiments, we noticed that other than records with an empty reference, datasets can contain also empty audio snippets. More precisely, these recordings do have a signal but it is mostly silence. We found this phenomenon primarily in Fleurs-es, with silence mostly coming from snippets attributed to women.

Therefore, we used `pyannote.audio`'s pretrained neural models and code for voice activity detection.[16] After counting the number of segments where voice was detected, we filter out all snippets were no segment was detected. We release the the counts and IDs of silence records in our repository.

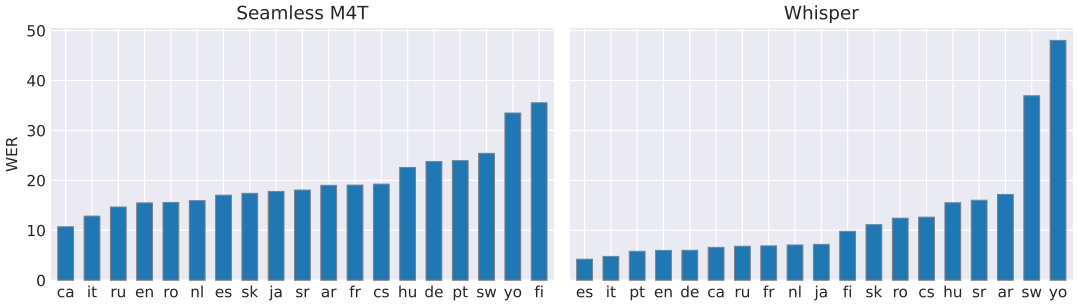
A.4 Speaker ID Attribution in `Fleurs`

Speaker IDs are crucial in our pipeline to avoid over- or under-estimating performance gaps due to overly present speakers. Since `Fleurs` does not provide this piece of information, we attribute it automatically.

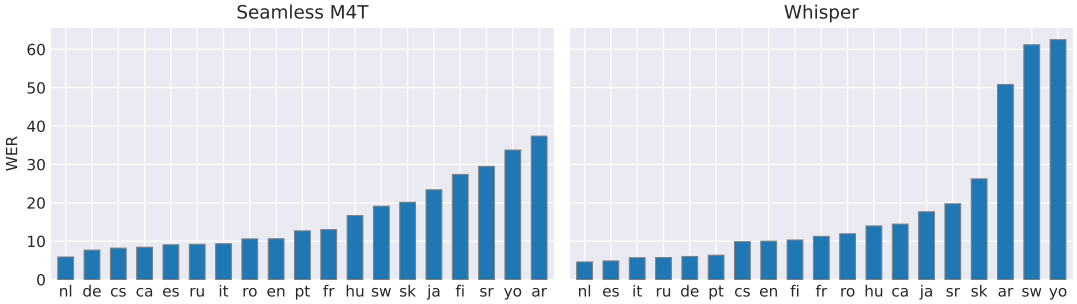
[16] https://github.com/pyannote/pyannote-audio



(a) VoxPopuli



(b) Fleurs



(c) CV

Figure 6: **SEAMLESS and WHISPER transcription quality.** Error rate results are computed on test splits.

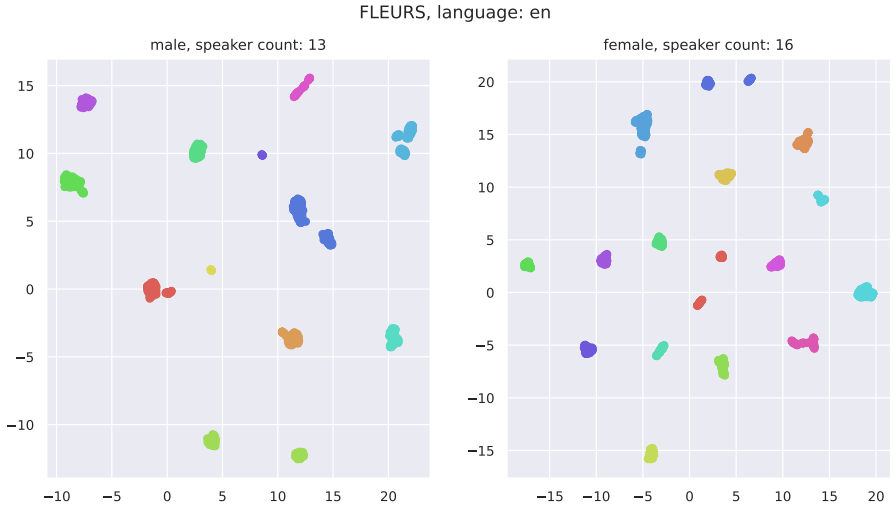


Figure 7: **UMAP projections of snippets in `Fleurs-en`.** We found 13 male speakers (left) and 16 female speakers (right). Color indicates cluster ID assigned by HDBSCAN.

		test	train	val	gender set	gender subset
Fleurs	Whisper	12.68±.11.3	12.64±.11.4	12.49±.11	12.60±.11.2	12.62±.11.2
	S M4T	19.87±.6.4	19.72±.6.3	18.75±.6.2	19.65±.6.3	19.67±.6.3
CV	Whisper	17.51±.17.5	16.65±.19.8	17.02±.18.7	16.88±.19.2	18.45±.19.2
	S M4T	16.42±.9.5	12.68±9.9	14.7±.8.9	13.06±.9.6	15.58±.9.3
VP	Whisper	13.59±.4.45	12.63±.3.79	12.99±.4.22	12.69±.3.8	13.32±.4.2
	S M4T	12.74±.3.27	11.27±.2.63	12.33±.2.68	11.39±.2.6	12.54±.2.9

Table 2: **Error rate distribution** across datasets splits (test/train/validation), all pre-sampled female/male records, and on the sampled gender subset we used. Results averaged over 19 languages in `CV` and `Fleurs`, and 11 for `VoxPopuli`.

		test	train	val	presample_all	overall
Fleurs	Whisper	-1.18±.2.9	-0.659±.1.8	-0.255±.2.6	-0.893±.1.7	-0.997±.1.9
	S M4T	0.153±.1.9	-0.880±.1	-0.257±.2.7	-0.735±.0.72	-0.809±.0.72
CV	Whisper	1.17±.5.7	-0.65±.2.7	0.88±.5.6	-0.14±.3.5	1.06±.5.3
	S M4T	2.01±.5.4	-1.22±.5.9	0.40±.2.7	0.59±.2.9	1.05±.3.3
VP	Whisper	0.31±.3.5	-0.80±.0.97	-0.58±.1.9	-0.73±.0.9	-0.104±.2.2
	S M4T	0.16±.2.8	-0.77±.0.54	-0.71±.2.1	-0.75±.0.6	-0.268±.2.2

Table 3: **Average F-M ER *absolute* difference** across datasets splits (test/train/validation), all pre-sampled female/male records, and on the sampled gender subset used in our experiments. The results are averaged over 19 languages for `CV` and 11 for `VoxPopuli`. For `Fleurs`, we average results over the only 9 languages comprising female/male speakers in all splits.

Language	**ISO**	**Family**
Japanese	ja	Japonic
Dutch	nl	Germanic
English	en	
German	de	
Swahili	sw	NC / Bantu
Yoruba	yo	NC / Volta-Niger
Catalan	ca	Romance
French	fr	
Italian	it	
Portuguese	pt	
Romanian	ro	
Spanish	es	
Arab	ar	Semitic
Czech	cs	Slavic
Russian	ru	
Serbian	sr	
Slovak	sk	
Finnish	fi	Uralic
Hungarian	hu	

Table 4: **Languages**, their ISO code, and family studied in this paper.

Specifically, we encode each snippet using a pre-trained ECAPA-TDNN model for speaker verification. We use code from SpeechBrain (Ravanelli et al., 2021) and the model checkpoint with Hub ID speechbrain/spkrec-ecapa-voxceleb. Next, we cluster embeddings separately per gender using HDBSCAN (Campello et al., 2013), with code from scikit-learn. We use set `min_cluster_size` to 2 and use cosine similarity. Finally, we assign a numerical ID to each found cluster. Figure 7 shows an example of clustered snippets projected with UMAP (McInnes et al., 2018). We release per dataset and language IDs in our repository.

A.5 Subset selection and data splits

To ensure the soundness of our experimental framework and of the gender subsample selection described in Section 3, we conducted preliminary analyses by comparing WER transcription quality scores (Table 2) and gender F-M gap results (Figure 3) across: *i)* different datasets splits (test/train/validation), *ii)* all female/male records available from any split, and *iii)* the sampled female/male subset used in our main experiments.

Since no precise information concerning the data

used to train WHISPER and SEAMLESS is publicly available, this evaluation is carried out with the goal of spotting potential data contamination issues. Namely, to ensure that we are not testing models on data samples comprised in their underlying training, thus posing the risk of obtaining unreliable results, which cannot be compared across datasets and models.

Table 2 shows no notable WER variations for Fleurs. Instead, both models perform better on the train splits of CV and VP, especially for the former dataset. The influence of the split accounted in CV and VP can thus be reflective of the F-M WER differences reported in Figure 3. Informed by such results, the gender subset used in our experiments is sampled from the test and validation split only for CV and VP, whereas for Fleurs, we leverage gender data from the whole corpus.

A.6 Dataset Statistics

Table 5, 7, and 6 reports all collected statistics on CV, Fleurs, and VoxPopuli, respectively.

A.7 Energy Statement

Experiments were conducted using a private infrastructure on hardware of type A100 PCIe 80GB (TDP of 250W) with carbon efficiency of 0.233 kgCO2eq/kWh.[17] Consumption is mostly due to inference to generate transcripts. We estimate a total emission of 3.6 kgCO2eq for experiments in the main body (transcripts of test and validation sets). Transcripts on training sets account for additional 104.24 kgCO2eq. No emission was directly offset. Total emissions are equivalent to driving an internal combustion engine car for 980 km.[18] Estimate was conducted using codecarbon[19] but does not account for energy for cooling the infrastructure.

B Details on the exploratory analysis

B.1 Acoustic Analysis

We provide further details about the acoustic analysis carried out in §5, comparing male and female speakers. Specifically, Table 9 reports the statistics of the T-tests conducted between acoustic features comparing genders across different languages and datasets, and Table 8 presents statistics on the Pearson correlation between differences in acoustic features and error rate gaps, separately for each (dataset, model) configuration.

We discuss the results of two representative languages only: Italian and Slovak. Italian exemplifies languages with comparable performance across genders, whereas Slovak is biased towards men and belongs to the set of languages with larger gender gap trends (see §4 for details). For both languages, acoustic features are, in most cases, statistically different. As shown in Figure 8, the most significant differences are found in mean pitch values, confirming the well-known differences between males and females. The distributions align with ranges suggested by the literature (Simpson, 2009) for both languages and all datasets. In contrast, differences in speaking rate are much less evident but still statistically significant according to the T-test (see Table 9). Intensity shows the most variability across all dimensions, even within gender groups, particularly noticeable in the Fleurs dataset.[20] In Italian (VoxPopuli), there is no difference in intensity between the two gender groups ($p = 0.356$), and similarly in Slovak (CV) ($p = 0.376$). However, for both languages, fitting an OLS regression at the sentence level using acoustic features to predict sentence-level error rates (r_F, r_M) showed no significant contribution, similar to other languages (R^2 with max: 0.20 and average$_\sigma$: $0.03_{\pm 0.04}$)

B.2 Lexical Analysis

To capture lexical phenomena, we extracted from the reference transcript of each record two sets of features. **Part-of-speech** tags include NOUN, PROPN, VERB, ADJ, ADV, PRON, AUX, CCONJ, DET, PART. **Named entities** tags include F_LOC, F_ORG, F_PER, and MISC. For each language, we used the corresponding SpaCy (https://spacy.io/models) "medium" text analyzer (e.g., for English en_core_web_sm and de_core_news_sm for German) when available. For hu an ar we used models available in Stanza (Qi et al., 2020). For yo and sw we limited the analysis to POS tags due to NE parses unavailability. We used mbeukman/xlm-roberta-base-finetuned-ner-yorub and mbeukman/xlm-roberta-base-finetuned-swahili-finetuned-ner-swahili for yo and sw, respectively (Beukman and Fokam, 2023). We left sr, sk, and cs out of this analysis since no taggers were available. We normalized

[17] https://app.electricitymaps.com

[18] Estimate based on average CO2 emissions of new passenger cars in EU in 2020. https://www.acea.auto/figure/average-co2-emissions-of-new-cars-in-eu

[19] https://github.com/mlco2/codecarbon

[20] This variability could be attributed to the underlying recording conditions of Fleurs, details of which were not provided in the original paper.

Lang	# records	Seconds	# Tokens	# M	Gini (M)	# F	Gini (F)	# O	Gini (O)
ar	14013	4.37±1.51	19.92±17.55	319	0.66	102	0.68		
ca	1645	6.12±1.73	17.05±6.22	265	0.22	241	0.22	8	0.26
cs	13746	4.60±1.38	20.77±8.71	293	0.64	39	0.61		
de	3641	6.10±1.75	16.74±6.08	807	0.22	129	0.24	17	0.24
en	4938	5.96±2.28	11.08±3.99	2262	0.25	476	0.25	33	0.21
es	6404	6.09±1.60	15.61±5.01	1476	0.29	517	0.30	42	0.34
fi	1695	4.74±1.66	18.22±8.93	52	0.57	17	0.57		
fr	4214	5.81±1.67	16.40±5.56	876	0.19	213	0.19	19	0.23
hu	12245	5.45±1.52	22.84±9.23	173	0.57	277	0.53		
it	7088	6.14±1.69	17.52±6.44	888	0.29	203	0.28		
ja	7981	5.02±2.29	22.64±15.07	1033	0.44	632	0.41		
nl	12808	4.83±1.38	16.29±5.57	487	0.62	130	0.63		
pt	7319	4.61±1.63	10.66±5.44	620	0.41	82	0.37		
ro	6489	4.05±0.90	15.70±4.13	167	0.56	39	0.57		
ru	10466	5.51±1.85	18.87±9.18	702	0.45	248	0.43		
sk	3008	4.34±1.74	13.66±9.64	58	0.64	13	0.65		
sr	1895	2.86±0.89	8.63±5.97	39	0.57	11	0.64		
sw	13730	5.61±1.80	21.04±8.13	248	0.62	251	0.58		
yo	1226	5.77±1.40	41.97±9.89	14	0.48	20	0.64		

Table 5: **Statistic for `CV`**. We used validation and test sets. Total number of records, average length in seconds, number of tokens as tokenized by Whisper's pretrained tokenizer, number of unique speakers and, gini index of the snippets-per-speaker dispersion for male (M), female (F), and "other" (O) subgroups.

Lang	# records	Seconds	# Tokens	# M	Gini (M)	# F	Gini (F)
cs	2208	9.93±5.38	60.04±31.51	45	0.58	16	0.47
de	4064	8.74±5.87	32.34±20.82	147	0.55	64	0.57
en	3485	9.92±6.17	24.48±15.05	274	0.52	104	0.52
es	3130	11.56±6.78	42.89±25.15	95	0.57	48	0.48
fi	1141	9.99±5.44	46.65±26.61	25	0.63	16	0.45
fr	3341	10.23±6.15	38.19±23.92	151	0.52	84	0.54
hu	2145	10.47±6.05	62.48±35.59	45	0.48	21	0.50
it	2419	13.00±7.19	52.78±28.99	87	0.47	37	0.53
nl	2351	8.16±5.23	36.63±22.88	69	0.54	42	0.54
ro	2701	11.29±5.75	57.31±31.54	51	0.52	19	0.57
sk	1266	10.33±5.55	62.52±35.16	26	0.57	10	0.47

Table 6: **Statistic for `VoxPopuli`**. We used validation and test sets. Total number of records, average length in seconds, number of tokens as tokenized by Whisper's pretrained tokenizer, number of unique speakers and, gini index of the snippets-per-speaker dispersion for male (M) and female (F) subgroups.

tags from different modes accordingly (e.g., we consider Stanza's PERSON and SpaCy F_PER the same feature). Moreover, we computed lexical complexity as specified in Imani and Habil (2017), i.e., the ratio between the sum of [NOUN, PROPN, VERB, ADJ, ADV] and the sum of [PRON, AUX, CCONJ, DET, PART].

B.3 Probing Analysis

Training and test set for probing experiments are sampled from the concatenation of each language's validation and test sets for consistency with error rate experiments. We balanced the samples on gender, stratifying on the speaker distribution.

We used standard logistic regression probes as per `scikit-learn`'s implementation and standard parameters but a larger allowance of steps for convergence (n=1000). For MDL probes, we used logistic regression as the backbone classifier for comparability and operationalized the probe via the *online code* configuration (Voita and Titov, 2020). Following previous work (Orgad et al., 2022), we set the dataset slices to (percentage): `[0.2, 0.4, 0.8, 1.6, 3.2, 6.25, 12.5, 25, 50, 100]`.

We probe a total of 1000 distinct positions that correspond roughly to the first tens seconds of each recording (Whisper encodes audio with 25 milliseconds-long rolling frame and stride of 10 milliseconds) (Radford et al., 2022).

C Release Statement

The following is a list of artifacts we produced in this work that we release to facilitate future research.

Lang	# records	Seconds	# Tokens	# M	Gini (M)	# F	Gini (F)
ar	836	10.83±4.08	57.52±21.70	9	0.31	38	0.51
ca	2914	11.87±4.00	39.93±14.87	5	0.15	5	0.05
cs	3772	12.05±4.45	53.38±19.93	7	0.36	5	0.21
de	3326	12.98±6.32	37.19±13.30	4	0.07	4	0.47
en	2832	9.75±3.58	24.55±8.54	13	0.33	15	0.28
es	3082	12.16±3.80	37.88±13.67	6	0.38	7	0.35
fi	3482	12.74±4.43	47.38±18.27	12	0.34	18	0.47
fr	3194	10.25±3.58	38.80±15.06	4	0.06	6	0.15
hu	4374	12.09±4.12	54.90±21.68	5	0.03	5	0.05
it	3528	14.53±5.34	40.59±14.30	6	0.33	6	0.31
ja	2348	12.99±3.77	49.50±18.25	6	0.50	16	0.66
nl	2594	9.55±3.73	40.59±14.63	8	0.53	6	0.23
pt	3944	12.50±4.24	35.32±12.88	5	0.08	4	0.19
ro	4022	10.22±3.63	48.98±18.03	5	0.08	7	0.35
ru	3496	11.39±3.97	42.57±16.14	13	0.36	11	0.26
sk	2600	11.65±3.81	53.15±19.93	8	0.41	10	0.57
sr	2820	10.76±3.40	59.46±21.59	4	0.07	6	0.08
sw	3588	14.08±4.52	49.60±17.25	18	0.42	17	0.39
yo	2850	16.32±5.66	73.91±34.05	7	0.36	6	0.50

Table 7: **Statistic for `Fleurs`.** We used validation and test sets. Total number of records, average length in seconds, number of tokens as tokenized by Whisper's pretrained tokenizer, number of unique speakers and, gini index of the snippets-per-speaker dispersion for male (M) and female (F) subgroups. Speakers are extracted automatically.

Dataset	Model	Feature	*rho*	*p*
`VoxPopuli`	SEAMLESS	Sp. rate	-0.10	0.76
		Intensity	-0.43	0.18
		Pitch	**-0.83**	**0.00**
	WHISPER	Sp. rate	-0.33	0.32
		Intensity	0.23	0.50
		Pitch	-0.30	0.36
`Fleurs`	SEAMLESS	Sp. rate	**0.62**	**0.00**
		Intensity	0.11	0.66
		Pitch	-0.06	0.82
	WHISPER	Sp. rate	0.20	0.42
		Intensity	-0.26	0.29
		Pitch	**-0.53**	**0.02**
`CV`	SEAMLESS	Sp. rate	**-0.51**	**0.02**
		Intensity	-0.25	0.31
		Pitch	0.20	0.41
	WHISPER	Sp. rate	-0.39	0.10
		Intensity	-0.10	0.69
		Pitch	0.38	0.11

Table 8: **Correlation between Acoustic Features and Error Rate Gaps**. Pearson correlation between the difference of the mean of subgroups an and $E(r_F, r_M)$. In bold the statistically relevant correlation ($p < 0.05$).

- Transcriptions obtained with Whisper and SeamlessM4T of the full datasets Mozilla Common Voice, Google Fleurs, and Meta VoxPopuli.
- Segment-level annotations of segments with voice activity extracted from the voice activity detection pipeline (§A.3).
- Segment-level acoustic features from our acoustic analysis (§5).
- Segment-level embeddings and cluster IDs extracted on Fleurs with the speaker identification pipeline (§A.4).
- Extensive statistics on speaker, gender, and record distributions on the three datasets—which we could not find anywhere else online.
- Dataset samples on which we computed the quality and gap metrics for reproducibility purposes (§3).

	ar	ca	cs	de	en	es	fi	fr	hu	it	ja	nl	pt	ro	ru	sk	sr	sw	yo
										Fleurs									
Intensity	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Pitch	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Speaking rate	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	✓
										VoxPopuli									
Intensity			✓	✓	×	✓	✓	✓	✓	×		✓		✓		✓			
Pitch			✓	✓	✓	✓	✓	✓	✓	✓		✓		✓		✓			
Speaking rate			✓	✓	✓	✓	×	×	✓	✓		×		✓		✓			
										CV									
Intensity	×	✓	✓	✓	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓	×	✓	✓	✓
Pitch	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Speaking rate	✓	×	✓	✓	×	×	✓	×	✓	✓	✓	✓	×	✓	✓	✓	✓	✓	✓

Table 9: **Statistical differences of acoustic features between genders**, separately by language and dataset. ✓indicates statistical difference, × indicates no difference (independent-sample Student's T-test, $p < 0.05$).

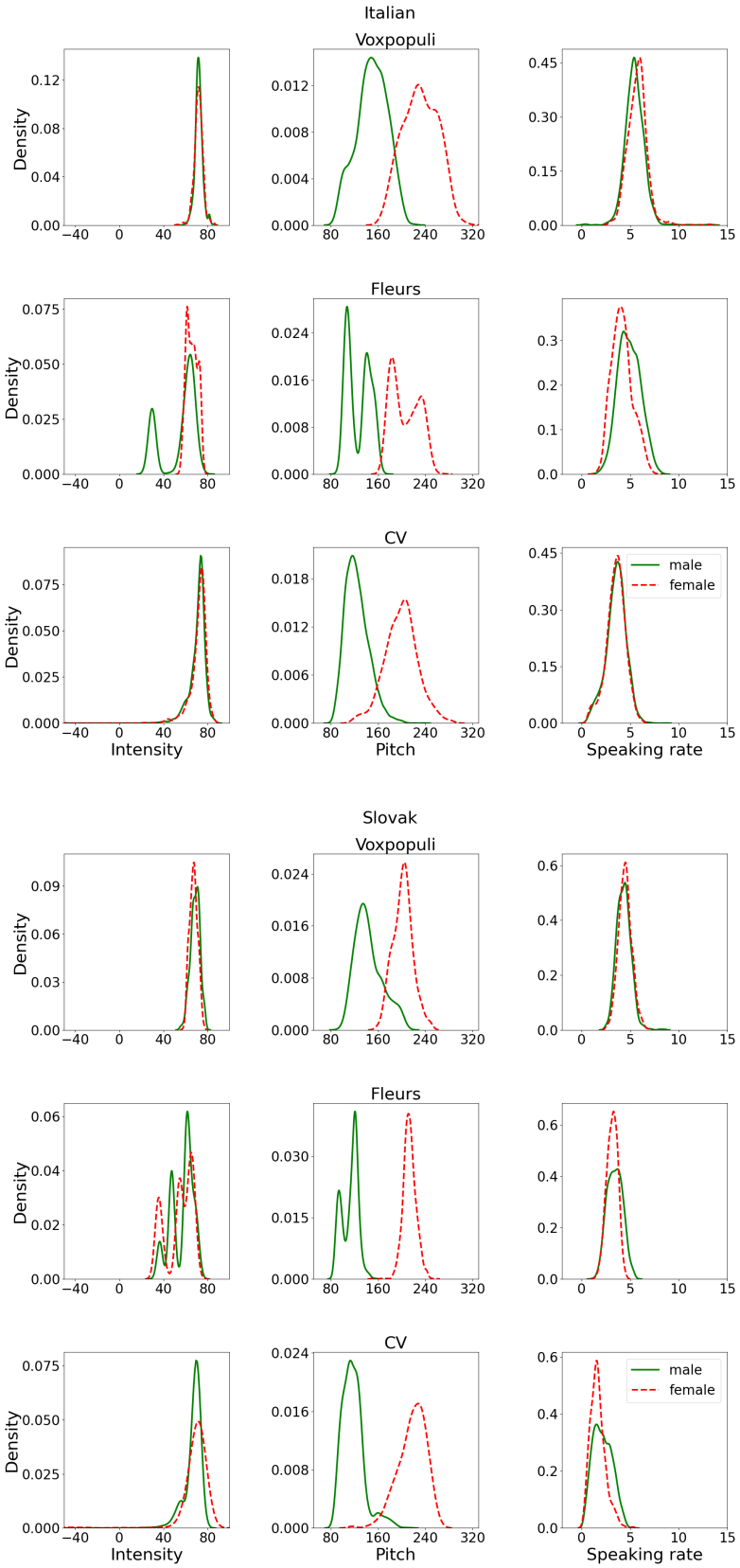
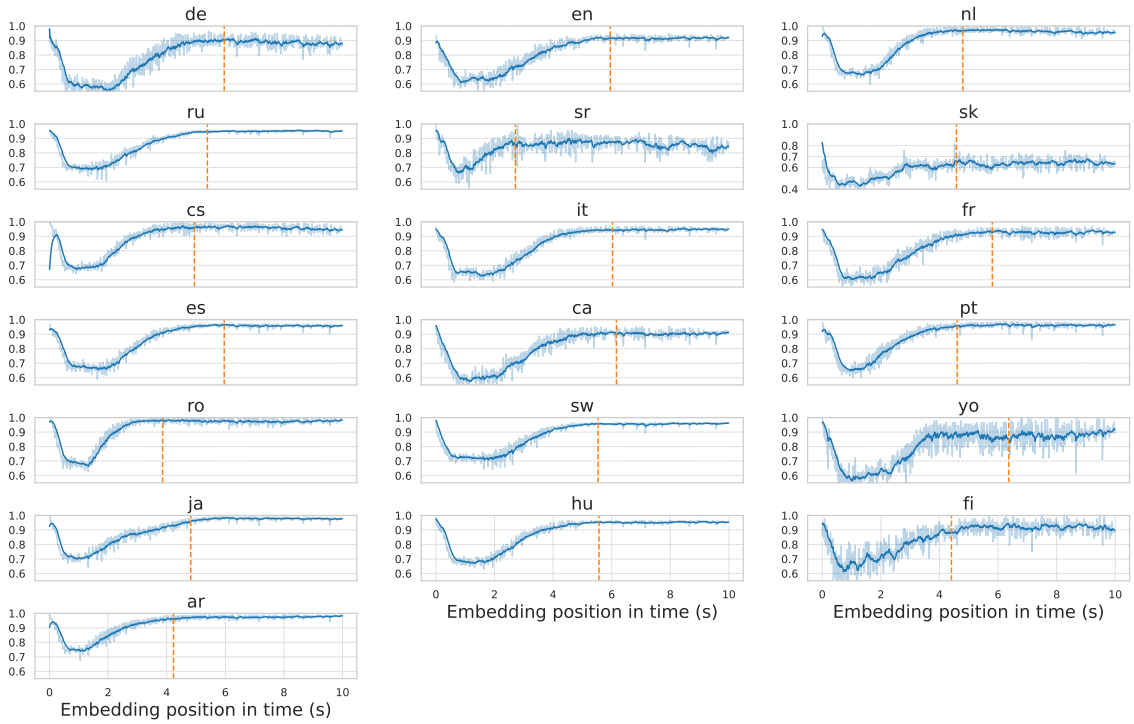
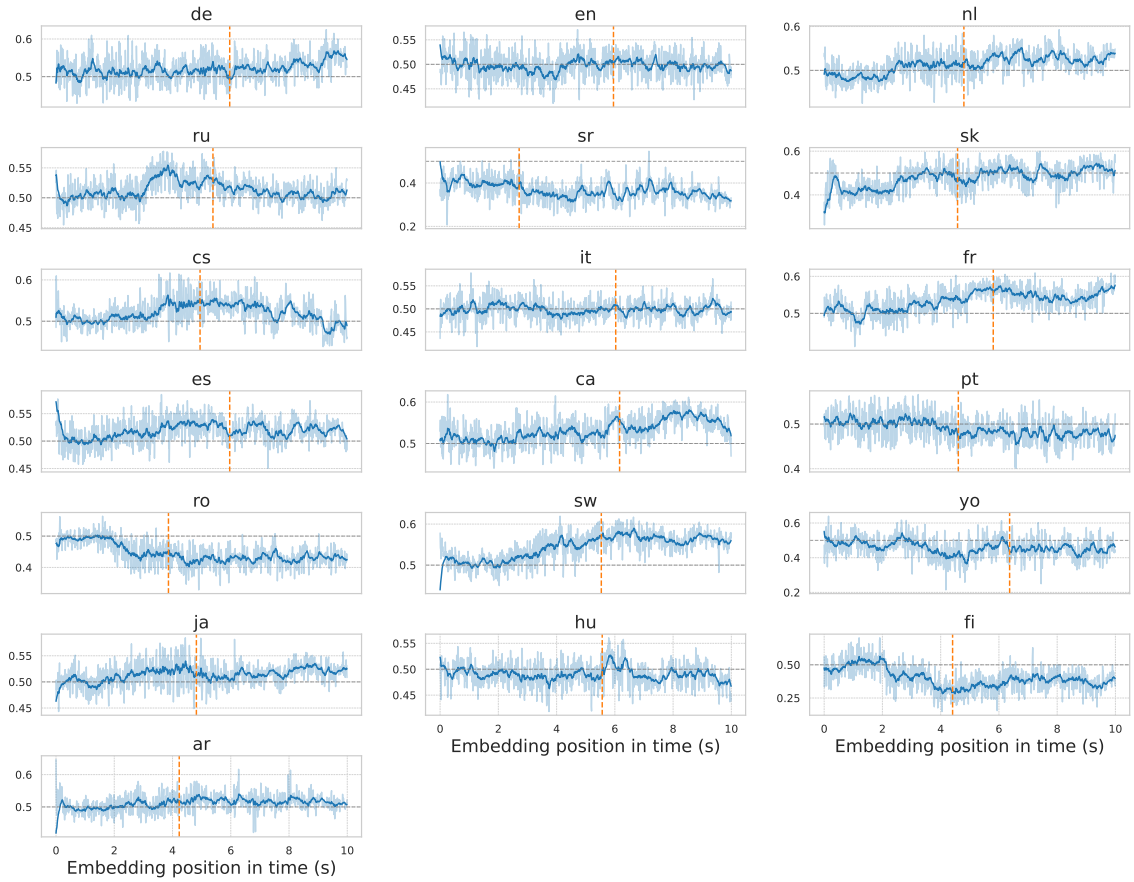


Figure 8: **Acoustic features.** Distributions of mean intensity, mean pitch, and speaking rate for Italian and Slovak on our three datasets. “Male” (green, solid line), “Female” (red, dashed line).



(a) Probes trained on the *original* labels.



(b) Probes trained on the *shuffled* labels.

Figure 9: F-M gender probing F1 Macro performance for every contextual embedding in Whisper-large-v3 within the first 10 seconds (x axis). Logistic regression probe with L2 regularization trained on standard (a) and shuffled (b) training labels. Orange lines indicate the average length of test segments. All CV languages.