Voice, Bias, and Coreference: An Interpretability Study of Gender in Speech Translation

Lina Conti[♠♦], **Dennis Fucci**[♠♦], **Marco Gaido**[♠], **Matteo Negri**[♠], **Guillaume Wisniewski**[♣], **Luisa Bentivogli**[♠]

[♠]Fondazione Bruno Kessler, Italy
{lvarellaconti,dfucci,mgaido,negri,bentivo}@fbk.eu

[♦]University of Trento, Italy

[♣]Laboratoire de Linguistique Formelle, Université Paris Cité, CNRS, Paris, France
guillaume.wisniewski@u-paris.fr

Abstract

Unlike text, speech conveys information about the speaker, such as gender, through acoustic cues like pitch. This gives rise to modality-specific bias concerns. For example, in speech translation (ST), when translating from languages with notional gender, such as English, into languages where gender-ambiguous terms referring to the speaker are assigned grammatical gender, the speaker's vocal characteristics may play a role in gender assignment. This risks misgendering speakers—whether through masculine defaults or vocal-based assumptions—yet how ST models make these decisions remains poorly understood. We investigate the mechanisms ST models use to assign gender to speaker-referring terms across three language pairs (en→es/fr/it). To do so, we examine how training data patterns, internal language model (ILM) biases, and acoustic information interact. We find that models do not simply replicate term-specific gender associations from training data, but learn broader patterns of masculine prevalence. While the ILM exhibits strong masculine bias, models can override these preferences based on acoustic input. Using contrastive feature attribution on spectrograms, we reveal that the model with higher gender accuracy relies on a previously unknown mechanism: using first-person pronouns to link gendered terms back to the speaker, accessing gender information distributed across the frequency spectrum rather than concentrated in pitch.

Keywords: gender bias, speech translation, interpretability, XAI

1. Introduction

Improved speech technology has made voice a popular modality to interact with AI systems, with applications like live translation through earphones now widely available (Chen, 2025). Unlike text, speech conveys information beyond linguistic content: vocal characteristics like pitch and pronunciation provide cues about the speaker's sociodemographic attributes, including gender, age, race, and social class (Labov, 1964; Thomas, 2010; Kraus et al., 2017; Simpson, 2009). This raises modality-specific concerns about social bias, as systems may perform differently across groups defined by these acoustic characteristics (Koenecke et al., 2020; Cercas Curry et al., 2024).

One example of these modality-specific concerns is the manifestation of gender bias when translating speech input compared to text input. When translating from languages with limited gender marking like English to languages with overt gender distinctions like Spanish, French, and Italian, systems assign grammatical gender to ambiguous terms. For example, when translating "*I have become a student*" to Italian, the verb form for "*become*" is grammatically gendered, leading the model to choose between "*diventata*"^{F} (feminine) and "*diventato*"^{M} (masculine). In text-based machine translation (MT), systems typically default to masculine forms or make assumptions based on gender stereotypes (Prates et al., 2018; Stanovsky et al., 2019; Mastromichalakis et al., 2025). Speech translation (ST) systems can exhibit similar patterns, but they also have access to vocal characteristics, such as pitch, that could be used as proxies for the speaker's gender when translating terms that refer to them (Bentivogli et al., 2020), as in the example above.

However, whether and how ST models use acoustic information for gender assignment remains poorly understood. While interpretability methods have been used to better understand gender bias in MT (Vanmassenhove et al., 2019; Wisniewski et al., 2022; Attanasio et al., 2023; Manna et al., 2025), research on the mechanisms underlying gender assignment in ST is scarce (Xu et al., 2023; Fucci et al., 2025; Yang et al., 2025). This gap is critical: without understanding how models make these ethically sensitive decisions, developing targeted mitigation for gender bias becomes

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 2232–2248
11–16 May 2026.

significantly more challenging.

To investigate the mechanisms ST models use to assign gender to speaker-referring terms, we start from the common assumption that attributes gender bias to training data imbalances (Tatman, 2017; Garnerin et al., 2019; Iluz et al., 2023; Mastromichalakis et al., 2025). This leads to our first research question: (i) **What is the influence of gender associations learned from the training data?** We address this by comparing model predictions with gender frequencies in the training corpus (§6). Finding that models do not simply replicate term-specific patterns motivates us to investigate the broader factors driving gender assignment in ST models. For this, we conceptually divide the ST model into two components: the encoder, which processes the input audio and may extract acoustic cues from it, and the decoder, which autoregressively predicts the next token based on both the encoder's representation of the audio and the previously generated tokens. First, we study the decoder's contribution by removing encoder information, thus isolating the ST system's internal language model (ILM) (Variani et al., 2020; Meng et al., 2021; Zeineldeen et al., 2021). Through this, we aim to answer the question (ii) **What is the impact of the model's learned knowledge of the target language and a priori assumptions about gender on predictions?** Following this analysis (§7), we turn to assessing the role of the source audio: (iii) **What aspects of the input audio does the model use to assign gender to speaker-referring terms?** Does it rely primarily on pitch, a key acoustic correlate of perceived gender? We study this through contrastive feature attribution over input spectrograms (§8).

The findings of our analysis across three language pairs (en→es/fr/it) challenge common assumptions about how ST systems perform gender assignment. The models we study do not simply replicate term-specific associations from training data, but learn broader patterns of masculine prevalence. While the ILM exhibits masculine bias, models can override these preferences. Moreover, they use first-person pronouns to link the gendered term back to the speaker, accessing vocal cues distributed across the frequency spectrum. This challenges the assumption that pitch would be the key acoustic feature (Bentivogli et al., 2020; Fucci et al., 2023a), as we find the first and second formants to be more important.

2. Bias Statement

Following Blodgett et al. (2020), we make explicit the assumptions underlying our work on bias. We focus on misgendering in ST: when systems translate speaker-referring terms into gendered target language forms that do not align with the speaker's gender identity. We consider outputs biased when they contradict reference translations that reflect the gender the speaker identifies with. When ST systems misgender speakers, allocational harms can arise through unequal performance: misgendered users must spend resources correcting system outputs (Savoldi et al., 2024). It also creates representational harms through the invisibilization of genders other than masculine, since models typically default to masculine forms. These harms affect women and gender non-conforming individuals. The binary gender framework we follow in this analysis does not allow us to cover the latter group; we discuss this limitation in §10.

3. Related Works

Gender bias in MT has been extensively studied (Savoldi et al., 2025a) with interpretability work revealing mechanisms underlying gendered choices in text-based systems. For instance, Wisniewski et al. (2022) and Manna et al. (2025) show that accurate gender disambiguation critically depends on correct handling of coreference chains. Feature attribution analyses (Attanasio et al., 2023; Sarti et al., 2023) further reveal that incorrect predictions typically result from models failing to attend to coreferring gendered pronouns and following a general masculine default rather than term-specific stereotypes. These insights from MT motivate our application of similar interpretability methods to ST, particularly given their success in informing mitigation strategies (Attanasio et al., 2023; Sarti et al., 2023).

However, ST introduces a distinct dimension that requires specific consideration. The same sentence sounds different depending on the speaker's gender, with acoustic variations in pitch, resonant frequencies, voice quality, and speech rate arising from both biological factors and sociocultural learned patterns (Simpson, 2009; Kreiman and Sidtis, 2011; Azul, 2015). Both humans and machines can detect and react to these variations (Tusing and Dillard, 2000; Chao and Bursten, 2021; Brown and Sonderegger, 2025). Studying whether and how ST models leverage these acoustic cues requires methods specifically designed for the speech modality, beyond those developed for gender bias in text-based systems.

Existing work on gender bias in speech technology has primarily focused on documenting performance disparities across demographic groups in various tasks: emotion recognition (Slaughter et al., 2023; Chien et al., 2024; Lin et al., 2024d), automatic speech recognition (ASR) (Adda-Decker and Lamel, 2005; Sawalha and Abu Shariah, 2013; Tatman, 2017; Garnerin et al., 2019; Feng et al., 2021;

Liu et al., 2022; Meng et al., 2022; Rajan et al., 2022; Attanasio et al., 2024), and ST (Zanon Boito et al., 2022; Costa-jussà et al., 2022; Bansal et al., 2025). Another common line of work studies the gender bias in speech technologies arising from what is said rather than how it sounds. Many studies adapt textual bias benchmarks by generating audio versions through text-to-speech systems (Lin et al., 2024a,c), which primarily test content-triggered bias. While this body of work establishes that gender bias is present in speech, it does not explain the mechanisms through which models use acoustic information.

Some recent studies have begun examining bias related to acoustic gender cues in spoken question answering (Choi et al., 2025a,b), finding that models largely fail to use them effectively and appropriately. But the translation task differs from question answering, and the pressure to assign grammatical gender could incentivize ST models to extract and use acoustic information differently.

Beyond measuring bias, some work has investigated the mechanisms behind gender bias in ST. Savoldi et al. (2022a) examined how it emerges during training, and Savoldi et al. (2022b) studied how tokenization choices affect gender bias patterns. However, these works do not explore how models use acoustic information for gender assignment, which is the focus of this paper.

Interpretability research has shown that speech representations encode gender information (de Seyssel et al., 2022; Chowdhury et al., 2024; Guillaume et al., 2024; Krishnan et al., 2024; Lin et al., 2024b; Fucci et al., 2025). However, these studies typically fail to establish causal links between encoded information and model outputs. We address this limitation by using perturbation-based feature attribution (§4.3) to identify features that causally drive gender assignment.

4. Method

To investigate the factors driving gender assignment by ST models for terms referring to the speaker, we examine three potential sources: training data patterns, the decoder's learned biases independent of the input audio, and the most relevant acoustic features from the input for gender assignment. In this section, we introduce the methods through which we investigate these aspects: comparison of the model's prediction with frequency patterns in the training data (§4.1); ILM analysis to examine the decoder's learned biases (§4.2); and contrastive feature attribution to identify the aspects of the audio driving gender assignment (§4.3).

4.1. Training Data Prevalence

To test the common assumption that gender bias is merely a direct reflection of training data distribution, we examine whether the model's gender preferences align with gender prevalence in its training data. The models we analyze (Wang et al., 2020; Papi et al., 2024) are trained on a single open-source dataset, MuST-C (Cattoni et al., 2021), which enables this analysis.

We identify gender terms referring to the speaker through string matching with the gender annotations in MuST-SHE (Bentivogli et al., 2020), our evaluation benchmark. For each such term in the translation hypotheses, we compute the prevalence of one gender g_1 (e.g., "*diventata*"$^{\text{F}}$) over the other g_2 (e.g., "*diventato*"$^{\text{M}}$) in the training corpus:

$$\text{Prevalence}(w_{g_1}, w_{g_2}) = \frac{\#w_{g_1}}{\#w_{g_1} + \#w_{g_2}} \quad (1)$$

where $\#w$ denotes the number of occurrences of word form w in the training data. We then compare the prevalence of term w in gender g_1 with the model's preference for that gender when generating w_{g_1}. We quantify this preference by computing the relative probabilities between gendered alternatives:

$$\text{Preference}(w_{g_1}, w_{g_2}) = \frac{p(w_{g_1})}{p(w_{g_1}) + p(w_{g_2})} \quad (2)$$

where $p(w)$ is the probability assigned by a given model to word w in the translation hypothesis. This can be calculated as either the predicted gender preference (comparing the generated form against its ungenerated alternative) or the masculine preference (comparing masculine versus feminine forms regardless of which was generated). By comparing prevalences with the model's gender preferences, we can distinguish predictions that replicate term-specific training patterns (where the higher-probability gender matches the more prevalent one in training data) from those based on acoustic or content-based source information, or other sources of bias.

4.2. ILM Approximation

While training data patterns provide one source of gender bias, the model's entrenched biases encompass more than only term-specific associations. The decoder's behavior also reflects its learned understanding of target language structure, general patterns of gender marking, and how previously generated tokens constrain subsequent predictions. As the decoder autoregressively predicts tokens based on both encoder input and previously generated tokens, it develops these broader linguistic patterns, forming an internal language model.

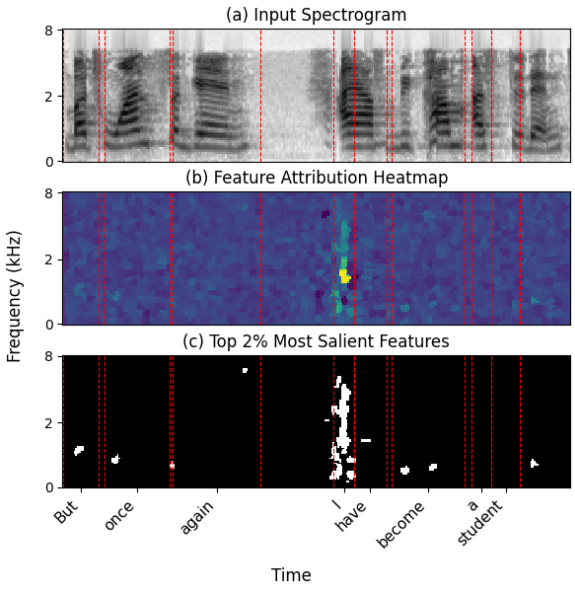


Figure 1: Example of contrastive feature attribution for the translation of "*become*" to "*diventata*"^{F} instead of "*diventato*"^{M}. (a) Input spectrogram. (b) Saliency heatmap showing features driving feminine gender assignment. (c) Top 2 % features sufficient to flip gender prediction.

To capture all aspects of these ingrained preferences that exist independently of the source audio, we analyze the ILM. Methods for its estimation have initially been developed for domain adaptation in ASR models (Variani et al., 2020; Meng et al., 2021; Zeineldeen et al., 2021). We adopt the ILM estimation method of Variani et al. (2020) and Meng et al. (2021), which replaces the encoder output with a dummy zero vector and has been shown to perform on par with more complex methods (Zeineldeen et al., 2021). While Fucci et al. (2023b) used the ILM to nudge ST models toward specific gender forms, here we analyze it to understand the decoder's inherent biases.

We compute the preference metric (Eq. 2) for the ILM and compare it with the full model's preferences. Contrasting ILM and full model preferences reveals to what extent models follow these entrenched biases and when, conversely, the audio input overrides them.

4.3. Feature Attribution

Understanding how models use input audio requires dedicated interpretability methods. Existing approaches for speech-to-text models (Trinh and Mandel, 2020; Markert et al., 2021; Mohebbi et al., 2023; Wu et al., 2023b; Fucci et al., 2024; Wu et al., 2024) primarily use perturbation techniques that mask input portions and measure the effect on model output. However, these methods generate holistic explanations that highlight features relevant for all aspects of word generation. Since our goal is to identify which input features drive gender assignment specifically, we employ the contrastive feature attribution method of Conti et al. (2025).

This approach identifies why the model generates one gendered form instead of its alternative by computing relative probability changes between the two options when different parts of the input are masked. Specifically, it automatically segments the input spectrogram (Figure 1.a) into acoustically meaningful regions and performs multiple inference passes with random perturbations to each segment. It assigns each segment a score based on how its perturbation affects the probability of one gendered form versus the other, producing saliency maps (Figure 1.b) that highlight the spectrogram regions most responsible for the model's gender choice.

Following Conti et al. (2025), we validate that the highlighted features are causally involved in gender assignment by testing whether occluding them changes the model's gender prediction. Figure 1.c shows the top 2 % of salient features that, when masked, successfully flip the prediction in this example. By occluding 1–20 % of the most salient features, we can flip the predicted gender in 37–47 % of examples across languages and models. We focus our analysis on such flipped examples, where causal links between input features and gender assignment are established. From these validated explanations, we can analyze which regions of the input spectrogram drive gender assignment: along the frequency dimension to identify relevant acoustic features, and along the time dimension to identify relevant words.

5. Experimental Setup

We describe below the data, models, and evaluation setup used throughout our analyses.[1]

Data We use MuST-SHE (Bentivogli et al., 2020), a benchmark containing annotations for gender-neutral English terms in natural speech that require gender marking when translated to Spanish, French, or Italian. We focus on the subset containing terms referring to the speaker, as these are cases where acoustic gender cues could influence gender assignment. Unlike sentences with gendered pronouns like "*She is a student*," where gender is explicitly marked in the source, speaker-referential sentences like "*I am a student*" contain no linguistic gender information, making acoustic cues potentially relevant. For each term, the

[1] Code to reproduce all analyses is available at https://github.com/lina-conti/voice-bias-coreference.

dataset provides correct and incorrect gender translations, e.g., "*diventata*"^{F} vs. "*diventato*"^{M} as Italian translations of "*become*," which we use as contrastive pairs for our analysis. Following Savoldi et al. (2022b), we exclude gender articles, as their high frequency in both genders across sentences makes it difficult to identify instances specifically referring to the speaker. We analyze only terms where the ST model generates one of the MuST-SHE annotated forms. Depending on the model used to translate, this yields between 309 and 453 gender terms per target language (see §11 for details).

Models We select models trained exclusively on a single open-source dataset to enable our training data analysis in §6. We therefore focus on two model architectures trained on MuST-C (Cattoni et al., 2021): the multilingual Transformer encoder-decoder model by Wang et al. (2020), and the monolingual Conformer encoder-Transformer decoder models from Papi et al. (2024). More recent ST systems and speech-enhanced large language models are typically trained on massive datasets that are not publicly released, making it difficult to establish connections between training data patterns and model behavior.

We primarily focus on the Transformer model for our analyses, as it demonstrates strong gender accuracy[2] for speaker-referential terms: 77.1 % to 80.6 % for feminine terms and 91.4 % to 94.4 % for masculine terms across target languages. This suggests that the Transformer is well positioned to leverage vocal cues for gender disambiguation, a phenomenon we aim to investigate further. By comparing these results with Conformer models, which achieve lower accuracy (39.2-49.8% for feminine terms and 72.5-76.7% for masculine terms across the three language pairs), we assess whether gender assignment strategies are model-dependent. These models provide architectural (Transformer vs. Conformer encoders) and scope (multilingual vs. monolingual) diversity, enabling us to examine how gender assignment strategies vary across different ST system configurations.

6. Training Data Analysis

This section addresses our research question "What is the influence of gender associations learned from the training data?" Specifically, we measure whether models replicate term-specific gender patterns from their training data by computing gender prevalence (§4.1) in MuST-C (Cattoni et al., 2021), the training corpus for the models we analyze.

The first thing we observe is that the training data shows a clear masculine skew for the gender terms we study. If we compute the average prevalence in the training data of the masculine form over the feminine one for all speaker-referring gendered terms in the translation hypotheses of the Transformer model, this average ranges from 0.68 to 0.71 depending on the languages (es: 0.68; fr: 0.71; it: 0.68), with nearly identical values for the Conformer models (es: 0.68; fr: 0,72; it: 0,68).

	More Freq.	Less Freq.
F	24	173
M	221	22

(a) Spanish

	More Freq.	Less Freq.
F	22	130
M	187	16

(b) French

	More Freq.	Less Freq.
F	12	140
M	192	13

(c) Italian

Table 1: Distribution of examples by predicted gender and whether the predicted gender is more or less prevalent in the training data for that term. Results for the Transformer model (Wang et al., 2020).

Table 1 shows that the Transformer model does not simply replicate training data patterns when assigning gender. The table categorizes each predicted gendered term by whether the predicted gender (F or M) is the more or less prevalent form in the training data for that specific term. If models followed the heuristic of generating each term in its most frequent training data gender, predictions should consistently fall in the "More Freq." column. Instead, the model frequently predicts genders that are less prevalent for that specific term: 87.8 % of feminine generated terms in Spanish (173 of 197), 85.5 % in French (130 of 152), and 92.1 % in Italian (140 of 152) correspond to the less prevalent form in the training data. For masculine predictions, the model does tend to predict the more prevalent form (221 of 243 in Spanish, 187 of 203 in French, 192 of 205 in Italian). However, given the overall masculine skew, this reflects the general pattern in the training data rather than term-specific memorization. The Conformers show a similar pattern (see Table 7): the distribution of predictions relative to prevalence resembles Table 1, although with slightly more predictions aligning with

[2]The proportion of correct gender realizations among terms where the model generates one of the MuST-SHE annotated forms (Gaido et al., 2020). Full results on gender accuracy for all models and language pairs are reported in Appendix 14.1.

the more prevalent form. Crucially, none of the models closely follow term-by-term gender associations from the training data.

In summary, our results challenge the assumption that gender bias in ST simply reflects the training data distribution (Tatman, 2017; Garnerin et al., 2019; Iluz et al., 2023; Mastromichalakis et al., 2025). Our findings align with Conti and Wisniewski (2023) and Elghazaly et al. (2025), suggesting that gender bias cannot be exclusively reduced to training data imbalances. The data's masculine skew clearly influences model behavior, but not through simple memorization—rather, models internalize broader biases that we investigate through the ILM.

7. Internal Language Model Analysis

While the training data analysis has revealed that models do not simply memorize term-specific associations, gender assignment must still be driven by some combination of learned decoder preferences and input audio features. We first investigate whether the decoder has internalized broader biases beyond individual term associations by addressing our second research question: "What is the impact of the model's learned knowledge of the target language and a priori assumptions about gender on predictions?" The ILM analysis isolates these entrenched preferences by removing encoder information, measuring what the decoder learned, independently of source audio. Comparing ILM predictions with full model predictions then reveals when and how the audio input overrides these biases.

The ILM reflects and amplifies the masculine skew observed in the training data. Averaging over all gender terms, the ILM's preference for masculine over feminine ranges from 0.74 to 0.81 for the Transformer, depending on the language—higher than the training data prevalence of 0.68–0.71. When we separate by the gender that is ultimately generated by the full model, the ILM's masculine preference is 0.85–0.88 for masculine predictions and 0.58–0.71 for feminine ones (always above 0.5, even when generating feminine forms; see Appendix 14.3 for full results). For the Conformer models, average masculine preference is 0.63–0.64 (0.71–0.74 for masculine predictions, 0.48–0.49 for feminine ones). While the Conformers' masculine preference drops just below 0.5 for feminine predictions, for masculine ones it remains well above 0.5, suggesting some masculine bias, though weaker than for the Transformer.

However, the full model frequently overrides these entrenched biases. In the example in Figure 1, the prevalence in training data for masculine "*diventato*" is 0.57, and the ILM preference is 0.85, yet the full model's preference for the predicted feminine form "*diventata*" is 0.99, illustrating how acoustic input can supersede learned biases and training data statistics. Table 2 shows this is common: the Transformer frequently predicts genders to which the ILM assigns lower probability, particularly for feminine predictions. Instead, the Conformers seem to rely more on their ILM: the Pearson correlation between ILM and full model masculine preference is strong for Conformers (es: 0.65; fr: 0.62; it: 0.62), but weak to moderate for the Transformer (es: 0.45; fr: 0.38; it: 0.47).

	Higher Prob.	Lower Prob.
F	52	145
M	225	18

(a) Spanish

	Higher Prob.	Lower Prob.
F	33	119
M	195	8

(b) French

	Higher Prob.	Lower Prob.
F	46	106
M	195	10

(c) Italian

Table 2: Distribution of examples by predicted gender and whether the ILM assigns higher or lower probability to the predicted gender compared to the alternative. Results for the Transformer model (Wang et al., 2020).

This section has shown that the ILM exhibits a masculine-as-norm bias across our models, but ST systems vary in how much they rely on these entrenched preferences versus input audio. The Transformer's ILM is strongly biased toward masculine, yet the full model frequently overrides these preferences based on acoustic information. The Conformers show weaker ILM masculine bias but rely more on it, struggling to leverage source audio effectively. This analysis demonstrates that ST models combine acoustic gender cues with language model preferences to varying degrees. These findings demonstrate that acoustic input can play a substantial role in gender assignment, motivating us to investigate which specific aspects of the audio our models exploit for this.

8. The Role of Input Audio

This section addresses our third research question: "What aspects of the input audio does the model use to assign gender to terms referring to the speaker?" For this, we apply the feature attribution method from §4.3. Occluding 1–20 % of the most salient features highlighted by the saliency map flips the predicted gender in 40.7 % of Spanish

examples, 46.8 % of French examples, and 37.0 % of Italian examples for the Transformer model, with comparable rates for the Conformers (es: 41.1 %; fr: 44.9 %; it: 43.0 %). For these flipped examples, we have a guarantee that the highlighted features are causally involved in gender assignment, since if we occlude them, the model's prediction changes. We analyze these saliency maps to identify which words and acoustic cues in the source audio influence gender assignment.

8.1. Frequency Dimension

Prior work assumes that, since pitch is strongly associated with perceived gender, ST models use it when correctly assigning gender to terms referring to the speaker (Bentivogli et al., 2020; Elaraby et al., 2018; Fucci et al., 2023a). We empirically test this assumption for the first time.

Pitch is the perceptual correlate of the fundamental frequency F_0 (Jurafsky and James, 2009): utterances with higher F_0 sound higher pitched and more feminine, whereas male speech typically has lower F_0 (Simpson, 2009). To determine whether the model relies on pitch, we examine the intensity with which the pitch region (80-350 Hz, where the fundamental frequency lies) is highlighted in our heatmaps. We aggregate each gender term's explanation by taking the max score for each frequency bin over the time dimension, then average across all gender terms.

Surprisingly, the pitch range does not show the highest scores, suggesting it is not the most important region driving the choice of gender to refer to the speaker for the models we study. Figure 2 shows the average score across the frequency range for the Transformer model on the en→it split, with the same pattern holding for other languages and for the Conformer models (see Figure 3). Instead, the formant range (350–2,500 Hz) displays the highest scores, with two peaks corresponding to the first and second formants (F_1 and F_2). These formants are important for identifying the word being uttered, especially vowels (Jurafsky and James, 2009), but their exact frequency also varies by speaker and, notably, depending on the speaker's gender (Simpson, 2009). This is visible in Figure 2: for masculine terms, saliency peaks appear at approximately 500–900 Hz and 1,200–1,600 Hz, while for feminine terms the peaks shift to approximately 800–1,100 Hz and 1,400–1,800 Hz, broadly consistent with the ranges of F_1 and F_2 observed in English vowels for male and female speakers respectively (Hillenbrand and Clark, 2009). This suggests that the assumption that ST models should leverage pitch information to disambiguate the gender of terms referring to the speaker does not fully correspond to our model's behavior.

Still, while not the dominant feature, we cannot exclude that pitch plays a role in the model's decision process. In the example in Figure 1, occluding the top 2 % of features with the highest scores flips the gender of the translation of "*become*" from feminine to masculine. These features are concentrated along the time axis but spread across most of the frequency range and, crucially, they include the pitch region. This pattern holds for all our samples: 99.9 % of examples that flip contain at least one feature in the pitch range among those occluded for flipping (99.8 % for the Conformers).

Overall, we can conclude that the information that the model uses is distributed across the frequency range rather than concentrated in pitch alone, with particular emphasis on F_1 and F_2. This has implications for interventions to mitigate gender bias or neutralize the gender of audio samples, suggesting that approaches acting solely on pitch would be insufficient. Moreover, since salient features for gender disambiguation are concentrated along the time dimension but distributed along the frequency dimension, this suggests that *when* acoustic cues occur may be particularly important. We now turn to analyzing which source words correspond to these temporally concentrated features.

8.2. Time Dimension

	'I'			Self-referential		
	es	fr	it	es	fr	it
Flip	16.9	21.7	25.8	23.7	28.0	35.1
All	25.4	28.4	33.3	32.5	35.2	42.0

Table 3: Percentage of examples where the top-scoring source word is "*I*" or a self-referential word, for examples that flip and for all examples. Results for the Transformer model (Wang et al., 2020).

Analyzing which source words drive gender assignment reveals a surprising pattern: models rely primarily on first-person pronouns and determiners that refer to the speaker. Manual investigation of the contrastive heatmaps has revealed that "*I*" is the word most frequently highlighted to explain gender choice. To validate this observation quantitatively, we extract word-level scores from the spectrogram heatmaps by using Gentle[3] to obtain each source word's time range and selecting the highest feature score within that range as the word's score. Table 3 shows that, for the Transformer model, "*I*" is the top-scoring word in 16.9–25.8 % of examples, depending on the language. Including other self-referential expressions ("*I*", "*I'd*", "*I've*", "*I'm*", "*my*", "*me*", "*myself*") raises these percentages to 23.7–35.1 %. These percentages are even higher when considering all examples in our dataset rather

[3] https://github.com/strob/gentle

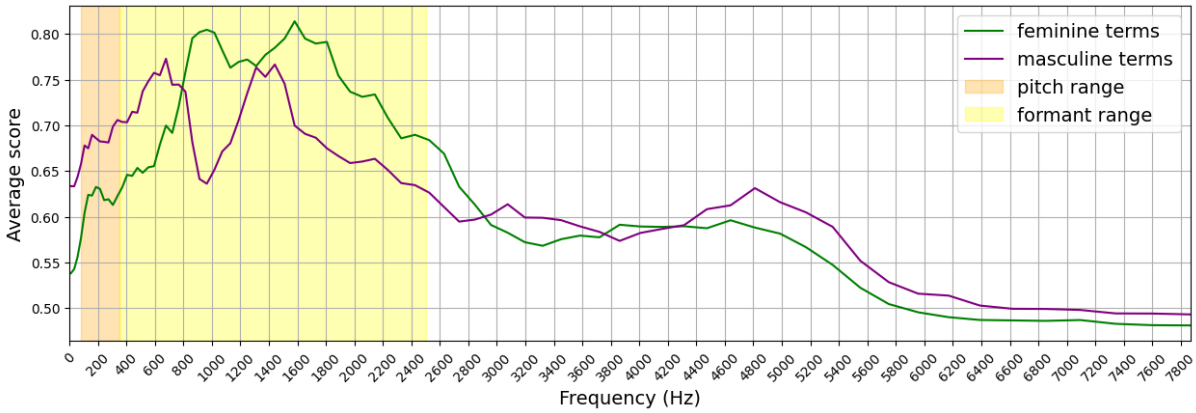


Figure 2: Average saliency scores per frequency bin (max-pooled over the time dimension, then averaged across all gender terms that flip), for the Transformer model (Wang et al., 2020) on en→it translation. Results are shown separately for feminine and masculine terms. Shaded regions mark the pitch range (80–350 Hz) and formant range (350–2500 Hz).

than just those that flip (second row of Table 3). Besides these first-person words, Table 4 shows that words like "*and*", "*was*", and "*when*" also frequently receive high attribution scores. However, manual inspection reveals these words appear next to "*I*" in the source sentence ("*and I...*", "*I was...*", "*when I...*") and likely score highest due to imprecisions in Gentle's word-level alignments, suggesting the actual prevalence of first-person words is even greater.

es		fr		it	
I	22	I	28	I	33
I'm	13	and	5	myself	9
sure	9	was	5	I'm	8
myself	9	I'm	4	as	7
as	8	musician	4	I've	6
scientist	7	me	3	scientist	6
kid	6	when	3	was	5
and	6	happy	2	and	4
child	6	serious	2	professor	3
us	4	lawyer	2	sure	3

Table 4: Most frequent top-scoring source words for examples that flip, with the number of examples for which they receive the highest saliency score in the source sentence. Results for the Transformer model (Wang et al., 2020).

The reliance on first-person pronouns differs across model architectures. For the Conformers, self-referential words are top-scoring in only 14.3–14.8 % of flipped examples (see Table 11), compared to 23.7–35.1 % for the Transformer. Despite this quantitative difference, the same set of words consistently appears at the top across all models and languages (Table 12), indicating that this strategy is learnable by different architectures but exploited with varying effectiveness. The Transformer, which achieves higher gender translation accuracy, relies on this mechanism more frequently than the Conformers. This pattern concurs with the finding of Fucci et al. (2025) that models with higher gender translation accuracy for speaker-referring terms also encode more gender information in their representations (measured via probing). Our contrastive analysis goes further by revealing the mechanism through which models access this information: via self-referential words that establish coreference with the speaker.

Beyond architectural differences, models' reliance on input features also differs between feminine and masculine predictions. For the Transformer, the percentage of examples whose gender prediction we can flip by occluding salient features in the input spectrogram is 60.8 % vs. 19.2 % for Spanish, 66.7 % vs. 24.9 % for French, and 52.2 % vs. 21.1 % for Italian (feminine vs. masculine respectively). The Conformers show the same pattern with a narrower gap (es: 47.5 % vs. 33.6 %; fr: 51.0 % vs. 37.6 %; it: 44.7 % vs. 41.3 %). This asymmetry suggests that the mechanism of accessing acoustic gender cues through first-person pronouns plays a more critical role for feminine predictions, while masculine predictions rely more heavily on the model's internal biases. This asymmetry aligns with prior work showing that language models and MT systems use masculine as a default, requiring strong feminine signals to generate feminine forms (Jumelet et al., 2019; Manna et al., 2025), which they slowly and imperfectly learn to use during training (Savoldi et al., 2022a).

This reliance on first-person pronouns reflects a mechanism analogous to coreference resolution in text-based MT. Just as MT models rely on gendered pronouns and determiners (e.g., "*she*"/"*he*", "*her*"/"*his*") that corefer with gender-ambiguous terms for disambiguation (Voita et al., 2018; Escudé Font and Costa-jussà, 2019; Stanovsky et al.,

2019; Manna et al., 2025), our ST models rely on first-person words that refer to the speaker. However, while gendered pronouns in text carry explicit gender information, words like "*I*" are semantically gender-neutral. In ST, acoustic gender cues in the speaker's voice effectively transform these semantically neutral words into gendered markers. Through coreference with the speaker, "*I*" provides access to the same gender information that explicit gendered pronouns provide in text-based MT, but encoded acoustically rather than lexically.

Besides first-person pronouns, models sometimes assign high salience to the source words corresponding to the gendered target terms. Table 4 contains examples of this phenomenon: words like "*scientist*", which translates to "*scienziato*"^{M} or "*scienziata*"^{F} in Italian. These words are directly relevant for translating the gendered term itself and, like all words in the utterance, carry acoustic gender cues that models can access.

In summary, the contrastive saliency maps reveal that ST models frequently rely on self-referential words like "*I*" to establish coreference with the speaker, enabling access to acoustic gender cues. These cues are distributed across the frequency spectrum rather than concentrated in pitch, with F_1 and F_2 showing higher importance.

9. Conclusion

We investigated how ST models assign gender to speaker-referring terms when translating from English to three Romance languages. Our analysis revealed that rather than memorizing individual term-gender pairings from training data, models learn that masculine forms are generally more prevalent. The decoder exhibits strong bias toward masculine defaults independent of input audio, but information from the audio can override these preferences. Crucially, models leverage first-person pronouns analogously to gendered pronouns in MT: acoustic cues transform the semantically neutral "I" into a functionally gendered marker, encoding gender information primarily in formant frequencies rather than pitch. These findings demonstrate that ST models can use vocal cues for gender disambiguation through mechanisms distinct from those assumed in prior work. They suggest that mitigation strategies centered on pitch manipulation (Fucci et al., 2023a) or exclusively rebalancing training data (Garnerin et al., 2019; Bansal et al., 2025) may prove insufficient, as they fail to account for the distributed nature of acoustic gender cues and how models mediate training statistics through complex learning dynamics.

10. Ethics Statement

Broader Impact. To mitigate harmful behaviors in AI systems such as those outlined in our Bias Statement (§2), we need both mitigation strategies (Vanmassenhove et al., 2018; Escudé Font and Costa-jussà, 2019; Saunders and Byrne, 2020; Saunders et al., 2022) and foundational research that reveals the mechanisms underlying biased behaviors. This interpretability work contributes to the latter effort, providing insights into how speech translation models make gender assignments that can inform future interventions.

Binary Gender Framework. Our analytical framework requires ethical consideration. We work within a binary gender framework, which offers methodological advantages—enabling contrastive explanations and leveraging existing benchmarks—but comes with significant limitations. This binary approach fails to account for gender identities outside the male/female binary (Zimman, 2020), potentially contributing to their erasure (Calado, 2025). Our contrastive analysis compares feminine versus masculine term generation without considering other possibilities such as gender-neutral reformulations (Piergentili et al., 2023; Savoldi et al., 2025b) or neologisms that avoid the binary dichotomy (Piergentili et al., 2024). While existing ST benchmarks like MuST-SHE (Bentivogli et al., 2020) provide binary gender annotations, comparable annotations for non-binary alternatives in speech do not yet exist. With such resources, we could potentially extend our analysis to include these forms, even though models rarely generate them spontaneously. We acknowledge that this resource-driven limitation means our analysis cannot capture the full spectrum of gender identities.

Gender Inference from Vocal Cues. Our analysis examines whether models use vocal cues for gender assignment, which could improve accuracy on binary benchmarks like MuST-SHE (Bentivogli et al., 2020). However, we do not advocate that models should rely on vocal characteristics to infer gender, as this risks equating gender with sex and treating it as biologically determined rather than as the behavioral and social phenomenon it is (Butler, 1990). Such an approach could harm speakers whose voices do not align with binary gender expectations. Instead, our goal is to understand what information current models exploit and how. Understanding these mechanisms provides a foundation for developing better control over model behavior, enabling different approaches depending on context and user preferences—whether that involves automatic inference, gender-neutral translations in ambiguous cases, or respecting user-specified pro-

nouns and gender identity.

11. Limitations

Models. Our analysis focuses on two model architectures trained on the MuST-C dataset (Cattoni et al., 2021). While these models are not state-of-the-art in terms of overall speech translation performance, we selected them for specific methodological reasons. The Transformer model (Wang et al., 2020) demonstrates notably higher accuracy on gender assignment compared to more recent systems like SeamlessM4T (Barrault et al., 2025), making it better positioned to reveal how models successfully leverage acoustic cues for gender disambiguation. Additionally, both models are trained exclusively on MuST-C, which enables the training data analysis in §6. In contrast, modern speech translation systems and speech-enhanced large language models are typically trained on multiple large-scale datasets without transparent documentation, making it difficult to establish connections between training data patterns and model behavior. The models we analyze provide architectural diversity (Transformer versus Conformer encoders) and scope differences (multilingual versus monolingual), allowing us to examine how gender assignment strategies vary across different system configurations. However, they are trained with the same objective (supervised training with cross entropy) and we do not cover different training regimes such as non-autoregressive or self-supervised trainings. Additionally, our analysis does not extend to speech-enhanced large language models (SpeechLLMs) (Rubenstein et al., 2023; Wu et al., 2023a; Tang et al., 2024), which represent an emerging paradigm for speech translation with potentially different mechanisms for processing acoustic information and assigning gender. Future work should investigate whether the patterns we identified—particularly the use of first-person pronouns for coreference resolution, the distribution of salient features across the frequency spectrum rather than concentration in pitch, and the relationship between internal language model biases and full model predictions—generalize to recent models trained on large-scale data, to models trained with different objectives, and to SpeechLLMs.

Language Coverage. Our analysis uses the MuST-SHE benchmark (Bentivogli et al., 2020), which covers the translation from English into three Romance languages (Spanish, French, and Italian). Since gender cues may manifest differently in other languages, this may influence how models learn gender assignment patterns. For this reason, future work should extend our analysis to typologically diverse language pairs, including languages with different gender marking strategies or non-gendered source languages. This would help determine whether the mechanisms we identify represent general model strategies or language-specific patterns.

Dataset Size. The benchmark also constrains our dataset size. After filtering as described in §5, we obtain 440 gender terms for Spanish, 355 for French, and 357 for Italian for the Transformer model, and 453 gender terms for Spanish, 379 for French, and 309 for Italian for the Conformer models. While this represents a relatively small dataset, MuST-SHE is currently the only available resource for speech translation that provides the fine-grained annotations required for our interpretability methods: gender terms referring to the speaker that are gender-neutral in English but gendered in the target language, gold-standard gender labels reflecting speakers' self-identified gender, and contrastive gender alternatives for each term. These annotations are essential for the contrastive feature attribution method we employ (Conti et al., 2025). The consistency of findings across three language pairs strengthens confidence in our results despite the dataset size.

12. Acknowledgments

This paper has received funding from the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU and from the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETings BetWEEN People).

13. Bibliographical References

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

Giuseppe Attanasio, Flor Miriam Plaza del Arco, Debora Nozza, and Anne Lauscher. 2023. A tale of pronouns: Interpretability informs gender bias mitigation for fairer instruction-tuned machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3996–4014, Singapore. Association for Computational Linguistics.

Giuseppe Attanasio, Beatrice Savoldi, Dennis Fucci, and Dirk Hovy. 2024. Twists, humps, and

pebbles: Multilingual speech recognition models exhibit gender performance gaps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21318–21340, Miami, Florida, USA. Association for Computational Linguistics.

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.

Shubham Bansal, Vikas Joshi, Harveen Chadha, Rupeshkumar Mehta, and Jinyu Li. 2025. Addressing speaker gender bias in large scale speech translation systems. *arXiv preprint arXiv:2501.05989*.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2025. Joint speech and text machine translation for up to 100 languages. *Nature*, 637(8046):587–593.

Luisa Bentivogli, Beatrice Savoldi, Matteo Negri, Mattia A. Di Gangi, Roldano Cattoni, and Marco Turchi. 2020. 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*, pages 6923–6933, Online. Association for Computational Linguistics.

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.

Jeanne Brown and Morgan Sonderegger. 2025. A sociophonetic study of creaky voice across language, gender and age in canadian english-french bilinguals. *Journal of Phonetics*, 112:101431.

Judith Butler. 1990. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge.

Filipa Calado. 2025. Some myths about bias: A queer studies reading of gender bias in NLP. In *Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 338–346, Vienna, Austria. Association for Computational Linguistics.

Roldano Cattoni, Mattia Antonino Di Gangi, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Must-c: A multilingual corpus for end-to-end speech translation. *Computer speech & language*, 66:101155.

Amanda Cercas Curry, Giuseppe Attanasio, Zeerak Talat, and Dirk Hovy. 2024. Classist tools: Social class correlates with performance in NLP. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12643–12655, Bangkok, Thailand. Association for Computational Linguistics.

Monika Chao and Julia RS Bursten. 2021. Girl talk: Understanding negative reactions to female vocal fry. *Hypatia*, 36(1):42–59.

Brian X. Chen. 2025. The new airpods can translate languages in your ears. this is profound. *The New York Times*.

Woan-Shiuan Chien, Shreya G. Upadhyay, and Chi-Chun Lee. 2024. Balancing speaker-rater fairness for gender-neutral speech emotion recognition. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 11861–11865.

Junhyuk Choi, Ro-hoon Oh, Jihwan Seol, and Bugeun Kim. 2025a. Voicebbq: Investigating effect of content and acoustics in social bias of spoken language model. *arXiv preprint arXiv:2509.21108*.

Junhyuk Choi, Jihwan Seol, Nayeon Kim, Chanhee Cho, EunBin Cho, and Bugeun Kim. 2025b. Acoustic-based gender differentiation in speech-aware language models. *arXiv preprint arXiv:2509.21125*.

Shammur Absar Chowdhury, Nadir Durrani, and Ahmed Ali. 2024. What do end-to-end speech models learn about speaker, language and channel information? a layer-wise and neuron-level analysis. *Computer Speech & Language*, 83:101539.

Lina Conti, Dennis Fucci, Marco Gaido, Matteo Negri, Guillaume Wisniewski, and Luisa Bentivogli. 2025. The unheard alternative: Contrastive explanations for speech-to-text models.

Lina Conti and Guillaume Wisniewski. 2023. Using artificial french data to understand the emergence of gender bias in transformer language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10362–10371.

Marta R. Costa-jussà, Christine Basta, and Gerard I. Gállego. 2022. Evaluating gender bias in

speech translation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2141–2147, Marseille, France. European Language Resources Association.

Maureen de Seyssel, Marvin Lavechin, Yossi Adi, Emmanuel Dupoux, and Guillaume Wisniewski. 2022. Probing phoneme, language and speaker information in unsupervised speech representations. In *Interspeech 2022-23rd INTERSPEECH Conference*.

Mostafa Elaraby, Ahmed Y Tawfik, Mahmoud Khaled, Hany Hassan, and Aly Osama. 2018. Gender aware spoken language translation applied to english-arabic. In *2018 2nd International Conference on Natural Language and Speech Processing (ICNLSP)*, pages 1–6. IEEE.

Hend Elghazaly, Bahman Mirheidari, Nafise Sadat Moosavi, and Heidi Christensen. 2025. Exploring gender disparities in automatic speech recognition technology. *CoRR*.

Joel Escudé Font and Marta R. Costa-jussà. 2019. Equalizing gender bias in neural machine translation with word embeddings techniques. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 147–154, Florence, Italy. Association for Computational Linguistics.

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

Dennis Fucci, Marco Gaido, Matteo Negri, Luisa Bentivogli, Andre Martins, and Giuseppe Attanasio. 2025. Different speech translation models encode and translate speaker gender differently. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1005–1019, Vienna, Austria. Association for Computational Linguistics.

Dennis Fucci, Marco Gaido, Matteo Negri, Mauro Cettolo, and Luisa Bentivogli. 2023a. No pitch left behind: Addressing gender unbalance in automatic speech recognition through pitch manipulation. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Dennis Fucci, Marco Gaido, Sara Papi, Mauro Cettolo, Matteo Negri, and Luisa Bentivogli. 2023b. Integrating language models into direct speech translation: An inference-time solution to control gender inflection. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11505–11517.

Dennis Fucci, Marco Gaido, Beatrice Savoldi, Matteo Negri, Mauro Cettolo, and Luisa Bentivogli. 2024. Spes: Spectrogram perturbation for explainable speech-to-text generation. *arXiv preprint arXiv:2411.01710*.

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.

Mahault Garnerin, Solange Rossato, and Laurent Besacier. 2019. 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*, pages 3–9.

Séverine Guillaume, Maxime Fily, Alexis Michaud, and Guillaume Wisniewski. 2024. Gender and language identification in multilingual models of speech: exploring the genericity and robustness of speech representations. In *Interspeech 2024*, pages 3330–3334. ISCA.

James M Hillenbrand and Michael J Clark. 2009. The role of f 0 and formant frequencies in distinguishing the voices of men and women. *Attention, Perception, & Psychophysics*, 71(5):1150–1166.

Bar Iluz, Tomasz Limisiewicz, Gabriel Stanovsky, and David Mareček. 2023. Exploring the impact of training data distribution and subword tokenization on gender bias in machine translation. 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 885–896, Nusa Dua, Bali. Association for Computational Linguistics.

Jaap Jumelet, Willem Zuidema, and Dieuwke Hupkes. 2019. Analysing neural language models: Contextual decomposition reveals default reasoning in number and gender assignment. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 1–11.

Dan Jurafsky and Martin H. James. 2009. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. Pearson Prentice Hall Upper Saddle River, NJ.

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.

Michael W Kraus, Jun Won Park, and Jacinth JX Tan. 2017. Signs of social class: The experience of economic inequality in everyday life. *Perspectives on Psychological Science*, 12(3):422–435.

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. In *Proc. Interspeech 2024*, pages 3090–3094.

William Labov. 1964. Phonological correlates of social stratification. *American Anthropologist*, 66(6):164–176.

Yi-Cheng Lin, Wei-Chih Chen, and Hung-Yi Lee. 2024a. Spoken stereoset: on evaluating social bias toward speaker in speech large language models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 871–878.

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

Yi-Cheng Lin, Tzu-Quan Lin, Chih-Kai Yang, Ke-Han Lu, Wei-Chih Chen, Chun-Yi Kuan, and Hung-yi Lee. 2024c. Listen and speak fairly: a study on semantic gender bias in speech integrated large language models. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 439–446. IEEE.

Yi-Cheng Lin, Haibin Wu, Huang-Cheng Chou, Chi-Chun Lee, and Hung-yi Lee. 2024d. Emo-bias: A large scale evaluation of social bias on speech emotion recognition. In *Proc. Interspeech 2024*, pages 4633–4637.

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. IEEE.

Chiara Manna, Afra Alishahi, Frédéric Blain, and Eva Vanmassenhove. 2025. Are we paying attention to her? investigating gender disambiguation and attention in machine translation. In *Proceedings of the 3rd Workshop on Gender-Inclusive Translation Technologies (GITT 2025)*, pages 1–16, Geneva, Switzerland. European Association for Machine Translation.

Karla Markert, Romain Parracone, Mykhailo Kulakov, Philip Sperl, Ching-Yu Kao, and Konstantin Böttinger. 2021. Visualizing automatic speech recognition–means for a better understanding? *ISCA Symposium on Security and Privacy in Speech Communication*.

Orfeas Menis Mastromichalakis, Giorgos Filandrianos, Maria Symeonaki, and Giorgos Stamou. 2025. Assumed identities: Quantifying gender bias in machine translation of gender-ambiguous occupational terms. *arXiv preprint arXiv:2503.04372*.

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

Zhong Meng, Sarangarajan Parthasarathy, Eric Sun, Yashesh Gaur, Naoyuki Kanda, Liang Lu, Xie Chen, Rui Zhao, Jinyu Li, and Yifan Gong. 2021. Internal language model estimation for domain-adaptive end-to-end speech recognition. In *2021 IEEE Spoken Language Technology Workshop (SLT)*, pages 243–250. IEEE.

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.

Sara Papi, Marco Gaido, Andrea Pilzer, and Matteo Negri. 2024. When good and reproducible results are a giant with feet of clay: The importance of software quality in nlp. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3657–3672.

Andrea Piergentili, Beatrice Savoldi, Dennis Fucci, Matteo Negri, and Luisa Bentivogli. 2023. Hi guys or hi folks? benchmarking gender-neutral machine translation with the gente corpus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14124–14140.

Andrea Piergentili, Beatrice Savoldi, Matteo Negri, and Luisa Bentivogli. 2024. Enhancing

gender-inclusive machine translation with neomorphemes and large language models. In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 300–314.

Marcelo O. R. Prates, Pedro H. C. Avelar, and L. Lamb. 2018. Assessing gender bias in machine translation: a case study with google translate. *Neural Computing and Applications*, 32:6363 – 6381.

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.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Gabriele Sarti, Nils Feldhus, Ludwig Sickert, and Oskar Van Der Wal. 2023. Inseq: An interpretability toolkit for sequence generation models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 421–435.

Danielle Saunders and B. Byrne. 2020. Reducing gender bias in neural machine translation as a domain adaptation problem. In *Annual Meeting of the Association for Computational Linguistics*.

Danielle Saunders, Rosie Sallis, and Bill Byrne. 2022. First the worst: Finding better gender translations during beam search. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3814–3823.

Beatrice Savoldi, Jasmijn Bastings, Luisa Bentivogli, and Eva Vanmassenhove. 2025a. A decade of gender bias in machine translation. *Patterns*, 6(6).

Beatrice Savoldi, Eleonora Cupin, Manjinder Thind, Anne Lauscher, Andrea Piergentili, Matteo Negri, and Luisa Bentivogli. 2025b. mgente: A multilingual resource for gender-neutral language and translation. *arXiv preprint arXiv:2501.09409*.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022a. On the dynamics of gender learning in speech translation. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 94–111. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022b. Under the morphosyntactic lens: A multifaceted evaluation of gender bias in speech translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1807–1824, Dublin, Ireland. Association for Computational Linguistics.

Beatrice Savoldi, Sara Papi, Matteo Negri, Ana Guerberof-Arenas, and Luisa Bentivogli. 2024. What the harm? quantifying the tangible impact of gender bias in machine translation with a human-centered study. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 18048–18076, Miami, Florida, USA. Association for Computational Linguistics.

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.

Isaac Slaughter, Craig Greenberg, Reva Schwartz, and Aylin Caliskan. 2023. Pre-trained speech processing models contain human-like biases that propagate to speech emotion recognition. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8967–8989, Singapore. Association for Computational Linguistics.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. Salmonn: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

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.

Erik R. Thomas. 2010. Teaching and learning guide for: Phonological and phonetic characteristics of african american vernacular english. *Lang. Linguistics Compass*, 4:737–741.

Viet Anh Trinh and Michael Mandel. 2020. Directly comparing the listening strategies of humans and machines. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:312–323.

Kyle James Tusing and James Price Dillard. 2000. The sounds of dominance. vocal precursors of perceived dominance during interpersonal influence. *Human Communication Research*, 26(1):148–171.

Eva Vanmassenhove, Christian Hardmeier, and Andy Way. 2018. Getting gender right in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3003–3008.

Eva Vanmassenhove, Dimitar Shterionov, and Andy Way. 2019. Lost in translation: Loss and decay of linguistic richness in machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 222–232.

Ehsan Variani, David Rybach, Cyril Allauzen, and Michael Riley. 2020. Hybrid autoregressive transducer (hat). In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6139–6143. IEEE.

Elena Voita, Pavel Serdyukov, Rico Sennrich, and Ivan Titov. 2018. Context-aware neural machine translation learns anaphora resolution. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1264–1274, Melbourne, Australia. Association for Computational Linguistics.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. Fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: System Demonstrations*, pages 33–39.

Guillaume Wisniewski, Lichao Zhu, Nicolas Ballier, and François Yvon. 2022. Analyzing gender translation errors to identify information flows between the encoder and decoder of a nmt system. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 153–163.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. 2023a. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Xiaoliang Wu, Peter Bell, and Ajitha Rajan. 2023b. Explanations for automatic speech recognition. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Xiaoliang Wu, Peter Bell, and Ajitha Rajan. 2024. Can we trust explainable ai methods on asr? an evaluation on phoneme recognition. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10296–10300. IEEE.

Chen Xu, Rong Ye, Qianqian Dong, Chengqi Zhao, Tom Ko, Mingxuan Wang, Tong Xiao, and Jingbo Zhu. 2023. Recent advances in direct speech-to-text translation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6796–6804.

Chih-Kai Yang, Neo S Ho, and Hung-yi Lee. 2025. Towards holistic evaluation of large audio-language models: A comprehensive survey. *arXiv preprint arXiv:2505.15957*.

Marcely Zanon Boito, Laurent Besacier, Natalia Tomashenko, and Yannick Estève. 2022. A study of gender impact in self-supervised models for speech-to-text systems. In *Proc. Interspeech 2022*, pages 1278–1282.

Mohammad Zeineldeen, Aleksandr Glushko, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2021. Investigating methods to improve language model integration for attention-based encoder-decoder asr models. In *Interspeech 2021*, pages 2856–2860.

Lal Zimman. 2020. Transgender language, transgender moment: Toward a trans linguistics. In *The Oxford Handbook of Language and Sexuality*. Oxford University Press.

14. Appendices

14.1. Model Gender Accuracy

Tables 5 and 6 report gender accuracy for all models and language pairs, complementing the model descriptions in Section 5. By gender accuracy, we

mean the proportion of correct gender realizations among terms where the model generates one of the MuST-SHE annotated forms (Gaido et al., 2020).

	es	fr	it
Feminine	80.60 %	77.35 %	77.09 %
Masculine	91.41 %	91.85 %	94.40 %

Table 5: Gender accuracy for the Transformer model (Wang et al., 2020).

	es	fr	it
Feminine	39.21 %	49.80 %	46.38 %
Masculine	76.74 %	72.50 %	75.76 %

Table 6: Gender accuracy for the Conformer models (Papi et al., 2024).

14.2. Training Data Analysis

Table 7 extends the results of Table 1 to the Conformer models.

	More Freq.	Less Freq.
F	22	116
M	284	31

(a) Spanish

	More Freq.	Less Freq.
F	19	131
M	213	16

(b) French

	More Freq.	Less Freq.
F	13	105
M	179	12

(c) Italian

Table 7: Distribution of examples by predicted gender and whether the predicted gender is more or less prevalent in the training data for that term. Results for the Conformer models (Papi et al., 2024).

14.3. Internal Language Model Analysis

Tables 8, 9, and 10 extend the results of Section 7 to all languages and models.

14.4. Feature Attribution

Tables 11 and 12 extend the results of Section 8 to the Conformer models, and Figure 3 extends the results in Figure 2 to all languages and models.

	es	fr	it
All terms	0.76	0.81	0.73
Fem. predictions	0.65	0.71	0.58
Masc. predictions	0.85	0.88	0.85

Table 8: ILM masculine preference for the Transformer model (Wang et al., 2020).

	es	fr	it
All terms	0.64	0.63	0.64
Fem. predictions	0.49	0.48	0.49
Masc. predictions	0.71	0.73	0.74

Table 9: ILM masculine preference for the Conformer models (Papi et al., 2024).

	Higher Prob.	Lower Prob.
F	72	66
M	228	87

(a) Spanish

	Higher Prob.	Lower Prob.
F	72	78
M	174	55

(b) French

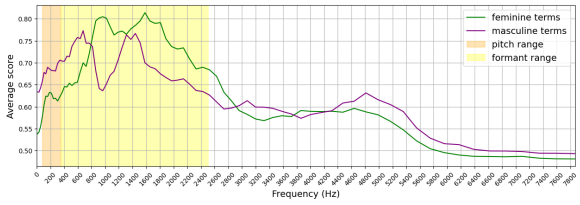
	Higher Prob.	Lower Prob.
F	56	62
M	148	43

(c) Italian

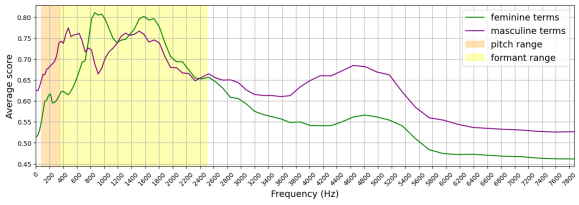
Table 10: Distribution of examples by predicted gender and whether the ILM assigns higher or lower probability to the alternative. Results for the Conformer models (Papi et al., 2024).

	'I'			Self-referential		
	es	fr	it	es	fr	it
Flip	10.1	9.0	15.5	14.3	12.1	14.8
All	8.6	6.7	20.1	11.6	10.0	20.4

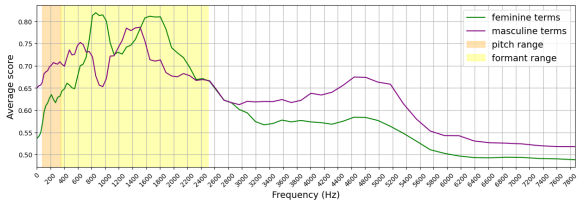
Table 11: Percentage of examples where the top-scoring source word is "I" or a self-referential word, for examples that flip and for all examples. Results for the Conformer models (Papi et al., 2024).



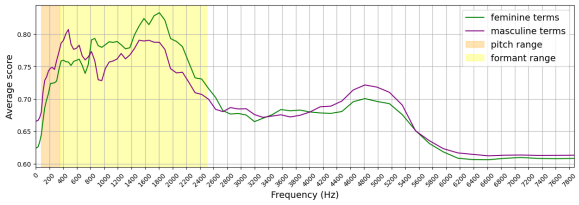
(a) Transformer, en→it



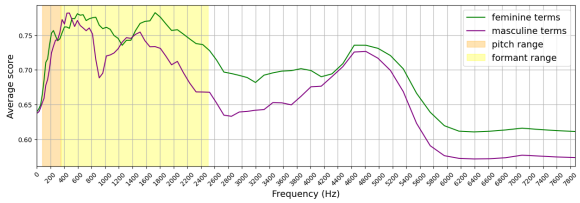
(b) Transformer, en→es



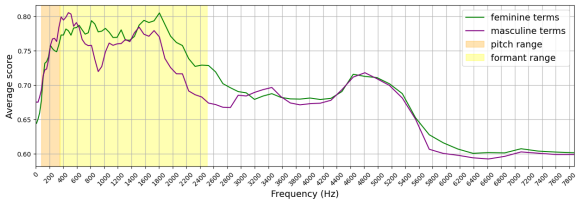
(c) Transformer, en→fr



(d) Conformer, en→it



(e) Conformer, en→es



(f) Conformer, en→fr

Figure 3: Average saliency scores per frequency bin (max-pooled over the time dimension, then averaged across all gender terms that flip), for the Transformer (Wang et al., 2020) and Conformer (Papi et al., 2024) models across all three language pairs. Results are shown separately for feminine and masculine terms. Shaded regions mark the pitch range (80–350 Hz) and formant range (350–2500 Hz).

es		fr		it	
scientist	12	sure	9	I	15
child	10	I	7	myself	5
I	10	went	5	scientist	5
kid	9	I'm	4	I'm	4
I'm	8	that	4	sure	4
as	5	ready	3	as	3
myself	5	love	3	brought	3
lawyer	4	and	3	been	3
us	3	when	3	child	3
a	3	researcher	3	student	3

Table 12: Most frequent top-scoring source words for examples that flip, with the number of examples for which they receive the highest saliency score in the source sentence. Results for the Conformer models (Papi et al., 2024).