Different Speech Translation Models
Encode and Translate Speaker Gender Differently

Dennis Fucci[αβ]**, Marco Gaido**[β]**, Matteo Negri**[β]**,**
Luisa Bentivogli[β]**, André F. T. Martins**[γδε]**, Giuseppe Attanasio**[ε]
[α]University of Trento, Italy [β]Fondazione Bruno Kessler, Italy
[γ]Unbabel, Portugal [δ]Instituto Superior Técnico, Portugal
[ε]Instituto de Telecomunicações, Portugal
dfucci@fbk.eu

Abstract

Recent studies on interpreting the hidden states of speech models have shown their ability to capture speaker-specific features, including gender. Does this finding also hold for speech translation (ST) models? If so, what are the implications for the speaker's gender assignment in translation? We address these questions from an interpretability perspective, using probing methods to assess gender encoding across diverse ST models. Results on three language directions (English → French/Italian/Spanish) indicate that while traditional encoder-decoder models capture gender information, newer architectures—integrating a speech encoder with a machine translation system via adapters—do not. We also demonstrate that low gender encoding capabilities result in systems' tendency toward a masculine default, a translation bias that is more pronounced in newer architectures.[1]

1 Introduction

Recent research in speech representation learning shows that models capture phonetic and speaker-related features in their internal representations (e.g., Prasad and Jyothi, 2020; Cormac English et al., 2022; Pasad et al., 2023; Yang et al., 2023; Choi et al., 2024; Shen et al., 2024; Chowdhury et al., 2024; Waheed et al., 2024). Building on sociophonetic studies that show how sociocultural and physiological factors shape voice production and perception (e.g., Coleman, 1976; Fuchs and Toda, 2010; Azul, 2013; Leung et al., 2018), recent research has explored how speaker *gender* is encoded in speech models. Studies have found evidence of gender encoding in self-supervised models (Chowdhury et al., 2024; de Seyssel et al., 2022; Guillaume et al., 2024) as well as in automatic speech recognition systems (Krishnan et al., 2024;

Attanasio et al., 2024). Despite growing interest, an analysis of gender encoding and its impact in speech translation (ST) remains largely absent.

Gender plays a crucial role in ST, particularly when translating from *notional* to *grammatical* gender languages. Here, models must infer gender from context and correctly apply inflections. Research suggests that encoder-decoder ST systems may use acoustic cues to assign grammatical gender to words referring to the speaker (e.g., En: "I was <u>born</u> in..." → Fr: "Je suis <u>né/née</u> à..."), yet masculine defaults remain common (Bentivogli et al., 2020; Gaido et al., 2020).[2] Moreover, as the ST field evolves, traditional encoder-decoder systems are being supplemented by newer architectures that integrate pretrained speech encoders into machine translation models via adapters (speech+MT). This raises key questions: ***whether* gender information is encoded and used by current ST systems**, and ***how* architectural variations influence this process**. Addressing these questions can shed light on gender-biased behaviors, such as the systematic preference for masculine defaults.

We investigate these aspects using probing, an established interpretability method (Conneau et al., 2018; Belinkov, 2022). We train probes to predict the speaker's gender[3] from hidden states in traditional encoder-decoder and speech+MT models. Then, by analyzing translations from English into Spanish, French, and Italian, we assess how gender encoding influences the translation of gender-marked terms referring to the speaker. Our results demonstrate that encoder-decoder models encode gender information, whereas speech+MT models encode it minimally, or do not encode it at all.

---

[1]Code available under Apache 2.0 at `https://github.com/hlt-mt/speech-translation-gender` .

[2]While inferring gender from acoustic cues is common—even among humans—it should not be treated as the default. This paper examines whether models exhibit similar behavior (for further discussion, see §8).

[3]We limit our analyses to binary gender, though we recognize that gender exists on a continuum (see §8).

Moreover, the ability to extract gender from hidden states correlates with how accurately models inflect gender in speaker-referred words. Strong probing performance is a proxy for high gender translation accuracy, while weak probing performance correlates with translations defaulting to masculine. These findings suggest that ST models do not uniformly rely on acoustic cues to translate speaker-referred expressions.

Our findings are compelling for several reasons. On the one hand, they show that encoding gender yields fairer systems for female users, challenging the conventional understanding in NLP that scrubbing gender information leads to fairer outcomes (Sun et al., 2019). On the other hand, relying on biometric markers for automatic decision-making contrasts broader and established ethical principles. We discuss the matter in §8.

2 Gender Probing

The encoder of a speech model maps an arbitrary audio signal into a sequence of d-dimensional hidden states $\mathbf{X} = \langle \mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L \rangle$, with $\mathbf{x}_l \in \mathbb{R}^d$ being the model's speech representation at position l and L the output sequence length. To evaluate whether and to which extent $\mathbf{X}$ encodes a specific attribute, classification *probes* (Belinkov, 2022) can be trained on states in $\mathbf{X}$, with classification accuracy indicating how well the attribute is encoded.

To obtain a single classification label for the whole sequence, previous works on gender encoding reduce its length L using mean or max pooling before training logistic probes (de Seyssel et al., 2022; Chowdhury et al., 2024; Guillaume et al., 2024; Krishnan et al., 2024) or non-linear probes (Krishnan et al., 2024). However, pooling may obscure positional variations in gender encoding, potentially weakening classification performance. Alternative approaches train separate probes on individual states $\mathbf{x}_l$ at specific positions l, assuming gender is not uniformly encoded across L. While more fine-grained than pooling, these methods have only considered a fixed, small set of relative positions (Krishnan et al., 2024) or do not support sequences of arbitrary length (Attanasio et al., 2024).

To avoid the limitations above, we design an attention-based probe. We draw inspiration from the Q-Former (Li et al., 2023), which maps a sequence of variable length into a fixed number of vectors by means of an attention mechanism where the query is a fixed sequence of learnable vectors. Similarly, our probe leverages a single learnable query to selectively extract gender information from hidden states. The sequence of hidden states $\mathbf{X}$ is projected into key ($\mathbf{K} \in \mathbb{R}^{L \times d}$) and value ($\mathbf{V} \in \mathbb{R}^{L \times d}$) matrices using learnable weight matrices $\mathbf{W}_K \in \mathbb{R}^{d \times d}$ and $\mathbf{W}_V \in \mathbb{R}^{d \times d}$, respectively. We then compute scaled-dot attention using $\mathbf{K}$, $\mathbf{V}$, and a learnable query vector $\mathbf{q} \in \mathbb{R}^d$. The output representation $\mathbf{o} \in \mathbb{R}^d$ is passed through a linear layer to compute class logits, and final probabilities are obtained via softmax.[4] The attention weights $\mathbf{a} \in \mathbb{R}^L$ instead indicate, for each input sequence, which hidden states contribute most to gender encoding. By using attention to selectively extract gender information from the entire input, followed by a single linear classification layer, the architecture remains simple—an important property for avoiding spurious correlations, as recommended for probing classifiers (Belinkov, 2022). At the same time, this attention-based design is more expressive than basic linear models (Hewitt and Liang, 2019), and serves as a proxy for how ST decoders access encoder states—namely, through the cross-attention mechanism. For a comparison with probes from previous works, see Appendix A.

3 Experimental Setup

Models We evaluate one ST encoder-decoder and two speech+MT models (full details in Appendix B) for en→es/fr/it translation. As encoder-decoder model (enc-dec), we use a Transformer-based model by Wang et al. (2020), trained on ST data from MuST-C (Cattoni et al., 2021). The speech+MT models are SeamlessM4T (Seamless—Communication et al., 2023) and ZeroSwot (ZeroSwot—Tsiamas et al., 2024). We probe the model states at different locations. In enc-dec, we use the encoder outputs. In speech+MT models, to assess the impact of adapters—which compress and map the speech encoder's outputs to the MT model's embedding space—we probe hidden states before (pre-ad) and after (post-ad) the adapters.

Data To train and evaluate probes, we use the MuST-C corpus, which includes speaker gender annotations based on self-declared pronouns (He/She) at the time of data collection (Gaido et al., 2020), thereby avoiding the risk of misgendering (see §8). For training, we randomly sample audio segments from the MuST-C training set (en→es section) to create gender-balanced training (train)

---

[4] See Appendix B for implementation details.

		test-generic	en→es			en→fr			en→it			Avg.
		All	She	He	All	She	He	All	She	He	All	All
Seamless	post-ad	59.76	90.41	24.30	51.72	86.18	29.55	54.51	85.98	32.14	55.63	53.95
	pre-ad	75.53	85.61	51.41	67.32	81.25	54.64	67.52	83.76	58.57	70.58	68.47
ZeroSwot	post-ad	59.57	85.98	38.73	61.80	90.13	37.80	61.62	86.35	39.64	60.65	61.36
	pre-ad	89.60	96.68	84.15	90.25	95.72	84.19	90.02	92.99	84.29	88.56	89.61
enc-dec		92.21	99.26	87.32	93.14	99.67	89.35	94.59	98.89	93.57	96.19	94.64

Table 1: Scores for gender probing (macro F1 for `All`, recall for `She`/`He`) across all configurations.

and validation (dev) sets. For testing, we use two datasets.[5] `Test-generic` consists of generic utterances taken from the validation and test sets of MuST-C. Since these sets are imbalanced across gender classes, `test-generic` is also inherently unbalanced. `Test-speaker` is drawn from MuST-SHE (Bentivogli et al., 2020), a corpus specifically designed to evaluate gender translation related to human referents from English into Spanish, French, and Italian. In MuST-SHE reference translations, each gender-marked word—corresponding to a neutral expression in English—is annotated with its opposite (wrong) gender form. We use the portion of MuST-SHE containing first-person references (e.g. En. "I was born" → Fr. "Je suis née$^\text{F}$<né$^\text{M}$>"), which is balanced across the two genders. Through `test-speaker`, we test gender encoding, gender translation, and their relationship, as the only cue for determining gender translation is the speaker's voice characteristics.

Evaluation We use macro F1 for overall probing performance, and recall for individual balanced classes. For overall ST quality, we use COMET (Rei et al., 2020).[6] For gender translation, we use the MuST-SHE evaluation script. It calculates *accuracy* as the percentage of speaker-referred words generated with the correct gender, out of all generated speaker-referred words. Instances where the model fails to predict the speaker-referred words—regardless of gender—are excluded from the accuracy calculation. The percentage of words included in the accuracy computation is reported as *coverage* (Gaido et al., 2020). Although out-of-coverage instances are excluded from gender accuracy computations, they can still convey gender information—for example, when the generated term is a gendered synonym. However, through

manual analysis of MuST-SHE translations by ST systems, Savoldi et al. (2022) demonstrated that out-of-coverage cases are typically either gendered synonyms that follow the same accuracy patterns as in-coverage examples, or mistranslations, where gender assessment is not applicable. We confirmed that this pattern also holds for the ST models investigated in this study through a manual analysis of out-of-coverage outputs, detailed in Appendix C.

4 Results

4.1 Gender Probing

Table 1 presents the macro F1 and single-class recall scores for the probes on `test-generic` and `test-speaker`. We first compare the results of probes tested on the final speech representations of `enc-dec` and `Seamless/ZeroSwot post-ad`. Overall, probes trained on the hidden states of `enc-dec` achieve high accuracy, with F1 scores peaking at 96.19 on `test-speaker` (`All` set, en→it) and scoring 92.21 on `test-generic`. In contrast, probes trained on `Seamless/ZeroSwot post-ad` yield significantly lower scores, not exceeding 59.76 on `test-generic` and 61.80 on `test-speaker` (en→es), respectively. Focusing on `test-speaker`, we note that `ZeroSwot post-ad` retains slightly more gender information than `Seamless post-ad`. Single-class recall scores are higher for `She` than `He` across all probes, especially in the weakest probe trained on `Seamless post-ad`. While this might suggest that the `He` class is harder to extract, the strong skew toward the `She` class in `Seamless` and `ZeroSwot post-ad` likely results from the probes' overall difficulty in extracting gender from these hidden states. Overall, these findings indicate that **speaker's gender encoding capability is high in the speech representations of the traditional encoder-decoder model, but significantly lower in speech+MT models**.

We now analyze the impact of the adapters in

---

[5]No speaker overlaps exist across `train`, `dev`, and `test` sets. For details on data statistics, see Appendix D.

[6]We use the `Unbabel/wmt22-comet-da` model.

		She			He			All		
		COMET	**Cov.**	**Acc.**	**COMET**	**Cov.**	**Acc.**	**COMET**	**Cov.**	**Acc.**
Seamless	**en→es**	80.81	70.37	12.09	81.69	70.56	90.55	81.26	70.47	53.62
	en→fr	78.31	59.61	17.60	79.50	59.66	88.84	78.89	59.63	53.29
	en→it	79.35	57.11	13.30	82.45	57.52	91.36	80.93	57.32	53.15
	avg.	79.49	62.36	14.33	81.21	42.61	90.25	80.36	62.47	53.35
ZeroSwot	**en→es**	85.57	78.04	51.50	83.56	73.72	76.73	84.54	75.79	64.46
	en→fr	82.99	65.52	54.38	82.20	64.55	70.11	82.60	65.03	62.20
	en→it	84.62	65.23	46.18	84.71	60.19	77.87	84.67	62.66	61.75
	avg.	84.39	69.60	50.69	83.49	66.15	74.90	83.94	67.83	62.80
enc-dec	**en→es**	77.03	66.93	81.47	75.64	65.69	90.94	76.32	66.29	86.45
	en→fr	71.77	53.20	77.68	73.68	53.79	91.34	72.70	53.50	84.62
	en→it	74.56	55.33	76.44	76.02	54.85	94.47	75.30	55.09	85.65
	avg.	74.45	58.49	78.53	75.11	58.11	92.25	74.77	58.29	85.57

Table 2: Translation performance in quality (COMET), gender coverage, and accuracy on `test-speaker` for all three language pairs, along with average scores.

Seamless and ZeroSwot. Looking at `pre-ad` results, `ZeroSwot` exhibits stronger probing performance compared to `Seamless` (avg. 89.61 vs 68.47 on `test-speaker`). This difference may stem from their training strategies: `Seamless` is jointly trained on speech and text, while `ZeroSwot` trains only the speech encoder, keeping the rest of the model frozen. Comparing `pre-ad` and `post-ad` results, we observe a substantial decrease in probing performance, with a $\sim 21\%$ drop in F1 for `Seamless` and $\sim 32\%$ for `ZeroSwot` after the adapters, across both test sets. These results indicate that **mapping speech representation into the MT embedding space via adapters significantly removes gender information in the speech+MT models.**

In summary, gender encoding varies across models. Speech+MT systems show lower encoding capability, particularly after the adapters, while the encoder-decoder model shows higher encoding capability, similarly to speech models explored in previous works. Additionally, we conducted an analysis to examine how gender information is distributed across the sequence length (see Appendix E). Interestingly, and consistent with prior findings on ASR models, we observe that ST models primarily encode gender in the early positions of the sequence.

4.2 Speaker's Gender Translation

Table 2 presents translation scores on `test-speaker` across the three ST models and language directions. Overall, `ZeroSwot` achieves the best translation quality, followed by `Seamless` and `enc-dec` (avg. COMET scores are respectively 83.94, 80.36, and 74.77). Coverage scores align with COMET trends, indicating that
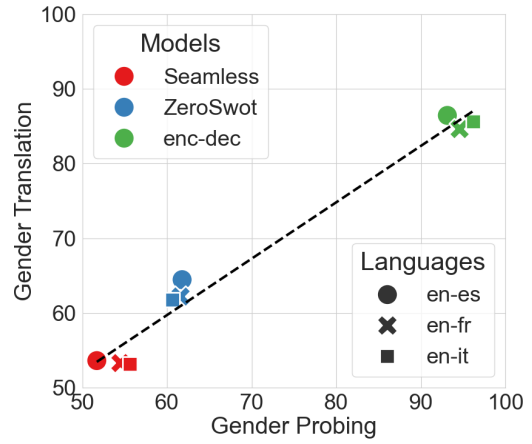

Figure 1: Correlation between overall gender probing performance (macro F1) and gender translation accuracy across models and languages on `test-speaker`.

higher translation quality increases the likelihood that system outputs include the speaker-referred words annotated in the reference. In all cases, more than half of the annotated words can be evaluated for gender translation.

Regarding gender translation accuracy, `enc-dec` provides the highest scores (avg. 85.57), while `ZeroSwot` and `Seamless` register lower average scores of 62.80 and 53.35, respectively. Notably, speech+MT models with stronger overall translation quality exhibit lower accuracy. Instead, as shown in Figure 1, accuracy scores strongly correlate with probing performance ($R^2 = 0.99$, $p < 0.01$). In other words, **higher gender encoding capability leads to more accurate gender translation**, supporting the idea that ST models rely on acoustic cues to translate speaker-referred

words when other signals are absent. When gender information is minimal, translations skew masculine. For instance, Seamless, for which gender encoding is almost absent (low F1 scores in Table 1), strongly favors masculine forms, with feminine score only peaking at 17.60 (en→fr). A similar bias is observed in ZeroSwot, where average accuracy is 50.69 for the She class and 74.90 for the He class. Even enc-dec exhibits a masculine skew: although it achieves higher feminine accuracy compared to the other models (avg. 78.53), this remains significantly lower than its masculine score (avg. 87.62), despite clear evidence of gender encoding (see Table 1). This suggests that an underlying linguistic bias sometimes overrides acoustic cues (see Appendix F for some examples).

All in all, ST models show variable performance on gender translation, with gender encoding serving as a proxy for the agreement between translation and the speaker's gender. When gender encoding is low, a masculine bias emerges.

5 Conclusion

We investigated how diverse ST architectures encode speaker gender information. Using attention-based probes, we found that while traditional encoder-decoder models trained solely on ST data retain this information, the adapters in newer speech+MT architectures tend to erase it. Moreover, gender encoding is correlated with the ST model's ability to assign the correct grammatical gender to words referring to the speaker. This finding suggests that ST models can leverage acoustic gender information when available. When such information is weakly encoded, the models default to masculine translations more frequently.

Our study sheds new light on how ST models encode and use gender information during translation, opening avenues for further research. For example, future work could explore which properties of adapters—or aspects of their training—contribute to the loss of gender information, and whether similar patterns emerge for acoustic and paralinguistic features that may be weakened during the mapping to the text embedding space. These findings also inform strategies to mitigate biases such as undue masculine translations. One approach could involve modifying the configuration of the adapters to better preserve the notion of the speaker's gender—when appropriate—or integrating external knowledge to avoid misgendering.

6 Acknowledgments

7 Limitations

Models and Languages Our analysis focuses on specific models and language directions. The choice of models was driven by their performance and widespread use in the selected language pairs. The selection of languages was based on the need to translate gender-neutral forms into gender-marked ones and the availability of annotated ST data. As more data becomes available, we plan to expand our analysis to include additional translation systems and languages, as well as systems which integrate speech into Large Language Models (Gaido et al., 2024). Nevertheless, prior research (Guillaume et al., 2024; Attanasio et al., 2024) suggests that gender encoding mechanisms are consistent across languages, so it is likely that similar patterns would emerge in other languages.

Probing While the probing paradigm (Conneau et al., 2018) is intuitive and well-established, it has been criticized for not directly confirming whether the model uses the extracted information (Belinkov, 2022). For example, some studies have highlighted a mismatch between probe performance and the original model's performance in NLP tasks (Belinkov and Glass, 2019; Ravichander et al., 2021; Elazar et al., 2021). Our experiments demonstrate a strong correlation between gender classification through probes (as an auxiliary task) and a specific translation aspect: gender assignment to words re-

ferring to the speaker (the original *sub*task). Although correlation does not imply causation, this finding proves that ST models likely use acoustic information for gender translation. To build on this, in future work, we plan to explore further interpretability techniques, such as *amnesic probing* (Elazar et al., 2021) or other methods in the field of mechanistic interpretability (Ferrando et al., 2024; Saphra and Wiegreffe, 2024).

8 Ethics Statement

Following prior work (de Seyssel et al., 2022; Guillaume et al., 2024; Chowdhury et al., 2024; Krishnan et al., 2024; Attanasio et al., 2024), we use the term "gender" as an approximation to account for documented differences in voices *and* linguistic expression of gender identity (She or He). Adopting this framework, we investigated how ST models encode gender-related vocal differences, which may stem from both physiological factors—such as anatomical differences in vocal tracts between male and female speakers (Simpson, 2001; Hillenbrand and Clark, 2009)—and sociocultural aspects—such as vocal behaviors associated with masculinity or femininity (Coleman, 1976; Nylén et al., 2024). Accordingly, to analyze how gender encoding capabilities affect ST systems' bias, we adopted a binary framework with only She/He class labels, primarily due to the lack of extensive speech data featuring non-binary voices. However, we recognize that vocal differences related to gender exist on a continuum, just as gender identities do. Expanding research to encompass non-binary identities is an essential next step.

In the context of translation, *gender bias* refers to the tendency of systems to favor one gender form over another—typically masculine over feminine—or to associate translations of specific roles and professions with a particular gender based on stereotypes (Savoldi et al., 2021). These biases can affect users' self-perception, as gendered language plays a fundamental role in shaping personal identity and representation (Stahlberg et al., 2007; Corbett, 2013; Gygax et al., 2019). As noted by Blodgett et al. (2020) and Savoldi et al. (2021), gender bias in translation technologies can lead to representational harms, such as reducing the visibility of women or reinforcing negative stereotypes about gender groups, as well as allocational harms, including disparities in the quality of service received by male and female users. Our experiments show that some models exhibit a strong bias toward masculine forms, posing a risk of such harms. We also find that encoding vocal characteristics related to gender can affect the accuracy of speaker gender translation.

Leveraging acoustic information was a key motivation behind the shift from cascaded systems—where vocal information was lost between ASR and MT—to direct models aimed at improving ST quality (Sperber and Paulik, 2020; Bentivogli et al., 2021). Our study provides a concrete example, demonstrating that accuracy of gender translation improves when gender-related vocal differences are encoded by the ST model. However, we do not necessarily advocate leveraging vocal features to assign linguistic gender markers, as such decisions may not align with a speaker's gender identity. This concern is particularly relevant for transgender individuals, children, and people with vocal impairments (Matar et al., 2016; Pereira et al., 2018; Villas-Bôas et al., 2021; Menezes et al., 2022). To avoid the risk of misgendering in our evaluation, we relied on self-declared gender identities. Although vocal properties in real-world scenarios may lead to misgendering and should be interpreted with caution, we argue that our findings nonetheless provide valuable insights. On one hand, our work contributes to advancing the understanding of gender bias in ST. On the other hand, raising awareness of how ST models encode speaker-related information—and how this influences translations—can help stakeholders make informed decisions to prevent harmful misgendering. It also empowers developers to make deliberate choices regarding model architecture and training strategies, as we found that specific architectures affect how gender information is encoded and used. Therefore, by enhancing the interpretability of ST models, we believe our work contributes to positive social impact.

References

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

David Azul. 2013. How do voices become gendered? a critical examination of everyday and medical con-

structions of the relationship between voice, sex, and gender identity. In *Challenging popular myths of sex, gender and biology*, pages 77–88. Springer.

Alexei Baevski, Henry Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Preprint*, arXiv:2006.11477.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

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

Luisa Bentivogli, Mauro Cettolo, Marco Gaido, Alina Karakanta, Alberto Martinelli, Matteo Negri, and Marco Turchi. 2021. Cascade versus Direct Speech Translation: Do the Differences Still Make a Difference? 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 2873–2887, Online. Association for Computational Linguistics.

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.

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.

Kwanghee Choi, Ankita Pasad, Tomohiko Nakamura, Satoru Fukayama, Karen Livescu, and Shinji Watanabe. 2024. Self-supervised speech representations are more phonetic than semantic. In *Interspeech 2024*, pages 4578–4582.

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.

Yu-An Chung, Yu Zhang, Wei Han, Chung-Cheng Chiu, James Qin, Ruoming Pang, and Yonghui Wu.

2021. W2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training. *Preprint*, arXiv:2108.06209.

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 Cora Meglioli, David Dale, et al. 2023. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation. *Preprint*, arXiv:2308.11596.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Greville G. Corbett. 2013. *The Expression of Gender*. De Gruyter.

Patrick Cormac English, John D. Kelleher, and Julie Carson-Berndsen. 2022. Domain-informed probing of wav2vec 2.0 embeddings for phonetic features. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 83–91, Seattle, Washington. Association for Computational Linguistics.

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*, pages 1402–1406.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. 2021. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-jussà. 2024. A Primer on the Inner Workings of Transformer-based Language Models. *Preprint*, arXiv:2405.00208.

Charlie Frogner, Chiyuan Zhang, Hossein Mobahi, Mauricio Araya-Polo, and Tomaso Poggio. 2015. Learning with a wasserstein loss. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'15, page 2053–2061, Cambridge, MA, USA. MIT Press.

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.

Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. 2024. Speech Translation with Speech Foundation Models and Large Language Models: What is There and What is Missing? *Preprint*, arXiv:2402.12025.

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.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 249–256, Chia Laguna Resort, Sardinia, Italy. PMLR.

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.

Pascal M. Gygax, Daniel Elmiger, Sandrine Zufferey, Alan Garnham, Sabine Sczesny, Friederike von Stockhausen, Lisa Braun, and Jane Oakhill. 2019. A language index of grammatical gender dimensions to study the impact of grammatical gender on the way we perceive women and men. *Frontiers in Psychology*, 10.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. 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.

Aravind Krishnan, Badr M. Abdullah, and Dietrich Klakow. 2024. On the encoding of gender in transformer-based asr representations. In *Interspeech 2024*, pages 3090–3094.

Yeptain Leung, Jennifer Oates, and Siew Pang Chan. 2018. Voice, articulation, and prosody contribute to listener perceptions of speaker gender: A systematic review and meta-analysis. *Journal of Speech, Language, and Hearing Research*, 61(2):266–297.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org.

Nayla Matar, Cristel Portes, Leonardo Lancia, Thierry Legou, and Fabienne Baider. 2016. Voice quality and gender stereotypes: A study on Lebanese women with Reinke's edema. *Journal of Speech, Language, and Hearing Research*, 59(6):1608–1617.

Danielle Pereira Menezes, Zulina Souza de Lira, Ana Nery Barbosa de Araújo, Anna Alice Figueirêdo de Almeida, Adriana de Oliveira Camargo Gomes, Bruno Teixeira Moraes, and Jonia Alves Lucena. 2022. Prosodic differences in the voices of transgender and cisgender women: Self-perception of voice - an auditory and acoustic analysis. *Journal of Voice*.

Frida Nylén, Johan Holmberg, and Maria Södersten. 2024. Acoustic cues to femininity and masculinity in spontaneous speech. *The Journal of the Acoustical Society of America*, 155(5):3090–3100.

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

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. Pytorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in python. *J. Mach. Learn. Res.*, 12(null):2825–2830.

Amanda Maria Pereira, Ana Paula Dassie-Leite, Eliane Cristina Pereira, Juliana Benthien Cavichiolo, Marcelo de Oliveira Rosa, and Elmar Allen Fugmann. 2018. Auditory perception of lay judges about gender identification of women with reinke's edema. *CoDAS*, 30(4).

Gabriel Peyré and Marco Cuturi. 2019. Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Archiki Prasad and Preethi Jyothi. 2020. How accents confound: Probing for accent information in end-to-end speech recognition systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3739–3753, Online. Association for Computational Linguistics.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377, Online. Association for Computational Linguistics.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Naomi Saphra and Sarah Wiegreffe. 2024. Mechanistic? In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 480–498, Miami, Florida, US. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9:845–874.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2022. 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.

Gaofei Shen, Michaela Watkins, Afra Alishahi, Arianna Bisazza, and Grzegorz Chrupała. 2024. Encoding of lexical tone in self-supervised models of spoken language. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4250–4261, Mexico City, Mexico. Association for Computational Linguistics.

Adrian P. Simpson. 2001. Dynamic consequences of differences in male and female vocal tract dimensions. *The Journal of the Acoustical Society of America*, 109(5 Pt 1):2153–2164.

Matthias Sperber and Matthias Paulik. 2020. Speech translation and the end-to-end promise: Taking stock of where we are. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7409–7421, Online. Association for Computational Linguistics.

Dagmar Stahlberg, Friederike Braun, Lisa Irmen, and Sabine Sczesny. 2007. Representation of the sexes in language. In Klaus Fiedler, editor, *Social Communication*, pages 163–187. Psychology Press.

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.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, et al. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

Ioannis Tsiamas, Gerard Gállego, José Fonollosa, and Marta Costa-jussà. 2024. Pushing the limits of zero-shot end-to-end speech translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14245–14267, Bangkok, Thailand. Association for Computational Linguistics.

Anna Paula Villas-Bôas, Karine Schwarz, Anna Martha Vaitses Fontanari, Angelo Brandelli Costa, Dhiordan Cardoso da Silva, Maiko Abel Schneider, Carla Aparecida Cielo, Poli Mara Spritzer, and Maria Inês Rodrigues Lobato. 2021. Acoustic measures of brazilian transgender women's voices: A case–control study. *Frontiers in Psychology*, 12.

Abdul Waheed, Hanin Atwany, Bhiksha Raj, and Rita Singh. 2024. What do speech foundation models not learn about speech? *Preprint*, arXiv:2410.12948.

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 2020 Conference of the Asian Chapter of the Association for Computational Linguistics (AACL): System Demonstrations*.

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.

Mu Yang, Ram C. M. C. Shekar, Okim Kang, and John H. L. Hansen. 2023. What can an accent identifier learn? probing phonetic and prosodic information in a wav2vec2-based accent identification model. In *Interspeech 2023*, pages 1923–1927.

Jinming Zhao, Hao Yang, Gholamreza Haffari, and Ehsan Shareghi. 2022. M-adapter: Modality adaptation for end-to-end speech-to-text translation. In *Interspeech 2022*, pages 111–115.

A   Probe Comparison

In our experiments, we use an attention-based probe, introduced in §2. It is designed to predict a

	Seamless pre-ad	ZeroSwot pre-ad	Enc-Dec	avg.
max pooling	61.37	54.46	65.48	60.44
mean pooling	62.17	76.44	90.10	76.24
positional sampling	71.45	64.96	83.37	73.26
attention-based	**75.53**	**89.60**	**92.21**	**85.78**

Table 3: F1 scores of probes on `test-generic` across different ST model states and probing strategies.

single label for each input sequence while preserving gender encoding across the temporal dimension. Unlike traditional approaches, our method avoids sampling or pooling mechanisms, which can undermine classification performance.

To assess its effectiveness, we compare our `attention-based` probe with three probing methods from prior work: logistic classifiers trained on hidden representations aggregated *i)* via max pooling, *ii)* via mean pooling, and *iii)* `positional sampling` of hidden states at relative positions. In the latter case, we sample hidden states at every 25% of the sequence length and train five separate probes, each corresponding to one of these positions. We train these probes on the hidden states extracted from Seamless/ZeroSwot pre-ad and enc-dec, as these representations encode the speaker's gender to a meaningful—although variable—extent. Table 3 reports the corresponding F1 scores on `test-generic`. For positional sampling, we report the highest scores among the five positions, which consistently correspond to the first position. Interestingly, this aligns with our analysis in Appendix E, where we show that attention weight distributions tend to concentrate at early positions.

Our results show that the `attention-based` probe consistently outperforms all other probing strategies across all considered settings. Notably, `max pooling` performs the worst (avg. 60.44), while `mean pooling` is the closest to the `attention-based` solution (avg. 76.24 vs. 85.78, respectively). The `positional sampling` method falls between `mean pooling` and `attention-based`, even though it surpasses `mean pooling` on `Seamless pre-ad`.

Overall, our approach proves to be more effective than traditional methods in our scenario. Moreover, our probe offers the additional advantage of interpretability by providing attention weights that reveal how gender information is distributed across the sequence (see Appendix E).

B Experimental Details

ST Models Below, we describe the architectures of the models used in our study (see §3), along with the details for running inference operations and training the probes.

- Transformer-based encoder-decoder model (Wang et al., 2020): it employs a convolutional downsampler to reduce the length of speech inputs, followed by a Transformer encoder—initialized from an ASR model— and a decoder. The encoder output is used for the `enc-dec` setting. Both the encoder and decoder are jointly trained for the ST task using autoregressive cross-entropy loss.

- SeamlessM4T (Communication et al., 2023): it uses a speech encoder based on the pre-trained w2v-BERT 2.0 (Chung et al., 2021), whose output is used in the `Seamless pre-ad` setting. A length adapter, derived from the M-adaptor (Zhao et al., 2022), processes the speech encoder's hidden states, and the resulting output is taken for the `Seamless post-ad` setting. The speech encoder followed by the length adapter is paired with a text encoder based on the pre-trained No Language Left Behind (NLLB) encoder (Team et al., 2022). Both encoders share a common text decoder, which is initialized from the NLLB decoder. The entire model is jointly trained using cross-entropy loss for speech-to-text and text-to-text translation tasks, along with token-level knowledge distillation from MT task (teacher) to the ST task (student).

- ZeroSwot (Tsiamas et al., 2024): it encodes speech using Wav2Vec 2.0 (Baevski et al., 2020), with its representations used in the `ZeroSwot pre-ad` setting. A CTC module then predicts characters from these representations. The CTC probabilities and Wav2Vec features are processed through an adapter to produce a compressed acoustic representation.

1014

This representation is then enhanced with positional encodings to form the final speech embedding, which is used in the `ZeroSwot` `post-ad` setting. The Wav2Vec encoder, CTC module, and adapter are fine-tuned by jointly minimizing the CTC loss and the Wasserstein loss (Frogner et al., 2015) using Optimal Transport (Peyré and Cuturi, 2019), to align speech-derived representations with the text-derived embeddings produced by the NLLB encoder.

Audio data are normalized to `float32` and truncated at 60s. Inference for all models is conducted using `transformers` `4.47.0` (Wolf et al., 2020) on an NVIDIA A40 GPU (48GB RAM). We use publicly available configuration files and model checkpoints from the HuggingFace Hub: https://huggingface.co/facebook/seamless-m4t-v2-large for SeamlessM4T, https://huggingface.co/johntsi/ZeroSwot-Large_asr-mustc_mt-mustc_en-to-8 for ZeroSwot, https://huggingface.co/facebook/s2t-medium-mustc-multilingual-st for the Transformer-based encoder-decoder.

Probes We implemented our attention-based probes using `PyTorch` `2.3.0` (Paszke et al., 2019). For training, we use a batch size of 32 and a starting learning rate of 0.0001, reduced by half if the validation loss does not decrease for 3 consecutive epochs. The weights of the classification layer are initialized using Xavier initialization (Glorot and Bengio, 2010). We use the Adam optimizer and optimize the model with cross-entropy loss. Training stops if the validation loss on the dev set does not improve by at least 0.00001 for 20 epochs. All training and testing were conducted on an NVIDIA Tesla K80 GPU (12GB RAM). For the probes used in Appendix A, we used logistic classifier implemented with `scikit-learn` `1.4.1` (Pedregosa et al., 2011) (`SGDClassifier` class with `log loss` and default values).

C Manual Evaluation

For the evaluation of gender translation, we adopt the official metrics from MuST-SHE (Bentivogli et al., 2020)—*gender accuracy* and *coverage* (see §3). Coverage is the percentage of sentences for which accuracy can be assessed. This number relies on automatic string matching, i.e., by checking whether the term annotated for gender appears, in any gendered form, in the system output. Accuracy is then measured only on the *in-coverage* (IC) outputs by verifying the gender marking of the word found. Therefore, coverage is complementary to accuracy. However, IC accuracy offers only a partial view, as *out-of-coverage* (OOC) instances— excluded from gender accuracy calculations—can still convey gender information, for example, when the generated term is a gendered synonym. To assess whether OOC instances exhibit trends that align with or diverge from the accuracy scores observed in IC cases, one of the authors manually analyzed the OOC outputs for Italian translations produced by the three ST models.

The manual analysis reveals that the OOC cases include: *(i)* translations that avoid speaker-referred gendered words by using paraphrases—which can sometimes be incorrect—or epicene terms; *(ii)* gendered synonyms; and *(iii)* gendered mistranslations referring to the speaker. Gender accuracy evaluation is possible only for cases of type *(ii)* and *(iii)*. Examples of all three categories are provided below.

(i)
En. source: *And to make sure* [. . .]
It. reference: *Per essere $\underline{sicuro}^{M}$ che* [. . .]
Seamless output: *E per $\underline{assicurarmi}$ che* [. . .]
Note: The model uses the verb "assicurarmi" (En. "to make sure") instead of the expression "essere sicuro" (En. "to be sure") found in the reference.

(ii)
En. source: [. . .] *I was amazed* [. . .]
It. reference: [. . .] *Ero $\underline{stupefatto}^{M}$.* [. . .]
ZeroSwot output: [. . .] *Ero $\underline{stupito}^{M}$.* [. . .]
Note: The adjective "stupito" is a synonym of "stupefatto", both meaning "amazed".

(iii)
En. source: *I'm exhausted, and I'm numb.*
It. reference: *Sono $\underline{esausta}^{F}$, e sono $\underline{paralizzata}^{F}$.*
Seamless output: *Sono $\underline{esausta}^{F}$ e sono $\underline{sordo}^{M}$.*

	She		He	
	Cov.	Acc.	Cov.	Acc.
String Matching				
Seamless	57.11	13.30	57.52	91.36
ZeroSwot	65.23	46.18	60.19	77.87
enc-dec	55.33	76.44	54.85	94.47
String Matching + Manual				
Seamless	70.56	12.59	73.79	93.75
ZeroSwot	78.43	45.95	76.94	80.76
enc-dec	71.57	78.55	70.39	95.32

Table 4: Coverage and accuracy scores for the She and He classes in the en→it direction. The upper three rows report results based solely on automatic string-matching evaluation (as in Table 2), while the lower three rows also include OOC instances where gender translation was manually evaluated.

> **Note**: The adjective "sordo" (En. "deaf") is a mistranslation of "numb" in this context.

Cases of type *(ii)* and *(iii)* account for 48.40% of the OOC instances for Seamless, 43.17% for ZeroSwot, and 64.64% for enc-dec. We incorporate these cases into the IC set and recompute the scores. Table 4 reports both the scores obtained using string-matching evaluation—as in Table 2—and those combining string matching with manual assessment of OOC instances. Coverage increases in the latter due to the reduced number of unassessable cases. More importantly, accuracy scores remain largely consistent. Absolute differences between string-matching scores and those including manual evaluation range from just 0.23 (for ZeroSwot, She class) to 2.89 (for ZeroSwot, He class). This consistency confirms that gender accuracy, when extended to include manually assessed OOC instances, aligns closely with that observed for IC cases, echoing findings by Savoldi et al. (2022) in their analysis of other systems.

Overall, this manual evaluation suggests that, despite the limitations of string-matching metrics, the trends they reveal remain broadly reliable, and the minor variations in gender accuracy introduced by manual evaluation do not affect the main findings.

D Data Statistics

Table 5 summarizes the distribution of samples, hours, and speakers across the dataset splits used for training and testing our probes.

The train set consists of 5061 samples, covering 9.70 hours of audio, with a balanced gen-

der representation across 1417 speakers. However, the number of speakers in the He class is higher than in the She class (883 vs. 534), reflecting the inherent imbalance in the MuST-C dataset (Cattoni et al., 2021). The dev set, approximately 20% the size of train, contains 1078 samples and 2.08 hours of gender-balanced audio from 337 speakers. The test-generic set includes 3702 samples spanning 6.69 hours, with a lower speaker-to-sample ratio than train and dev. Unlike these sets, test-generic is not gender-balanced, as it was created by merging the original test and validation sets of MuST-C, where male-speaker samples are significantly more numerous. Finally, the test-speaker set is the smallest, containing over 550 samples and 1.30 hours of audio, with small variations across language directions. Unlike test-generic, this set maintains gender balance.

E Gender Encoding Across Sequence Lengths

For each input sequence, our probes provide attention weights **a** that indicate the positions in the sequence where the probe primarily focuses when predicting gender (see §2). These trends can serve as a proxy for understanding how gender information is distributed across the sequence length, particularly for probes that effectively extract gender-related features. To analyze the overall weight distribution, we compute the average attention weight distribution over all test instances. However, since hidden state sequences vary in length depending on input audio duration, we interpolate the attention weight sequences to a fixed length of 100 elements. Figure 1 shows the average distribution of these resampled attention weights for all considered probes.

In probes trained on the enc-dec, Seamless pre-ad, and ZeroSwot pre-ad/post-ad representations, we observe an initial peak in attention followed by a descending trend. However, Seamless pre-ad shows a sharp peak at the very beginning of the sequence, suggesting that only the earliest portion is primarily used for classification. In contrast, for Seamless post-ad, where the probe performs worst, the initial positions appear to be the least relevant, although no clear pattern emerges. This could be due to the absence of gender encoding, leading to suboptimal probe performance.

Overall, our findings suggest that the earliest

	# Samples			# Hours			# Speakers		
	All	**She**	**He**	**All**	**She**	**He**	**All**	**She**	**He**
train	5061	2514	2546	9.70	5.08	4.62	1417	534	883
dev	1078	539	539	2.08	1.10	0.98	337	132	205
test-generic	3702	1026	2676	6.69	1.89	4.80	38	11	27
test-speaker (en→es)	555	271	284	1.33	0.69	0.64	202	104	98
test-speaker (en→fr)	595	304	291	1.36	0.71	0.65	189	88	101
test-speaker (en→it)	551	271	280	1.31	0.68	0.63	147	69	78

Table 5: Statistics of number of samples, hours, and speaker for the data splits used.

(a) test-generic   (b) test-speaker (average across language pairs)
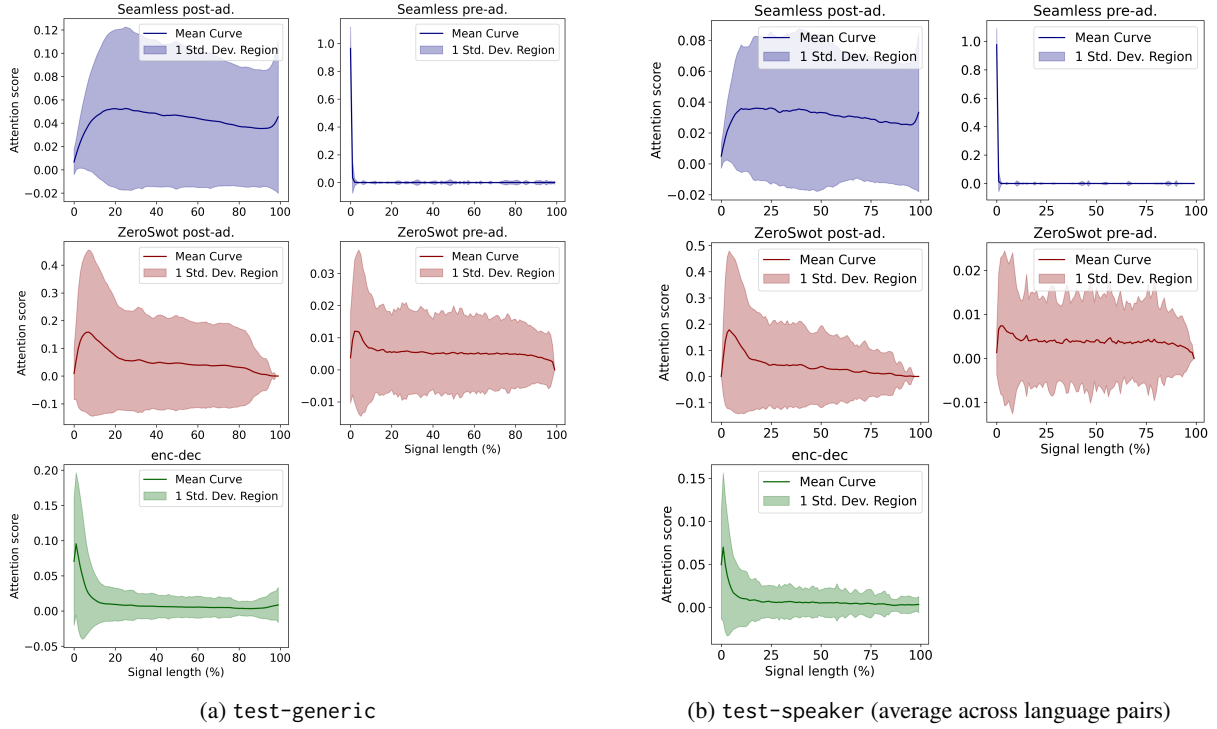
Figure 2: Mean attention weights with standard deviations for the various attention-based probes.

positions in the sequence are particularly important for predicting gender, especially when the probe is highly effective. This aligns with previous work by Krishnan et al. (2024) and Attanasio et al. (2024), who found that gender encoding is predominantly concentrated at the beginning of the sequence. However, it is important to note that the specific positions attended to can vary depending on the probe, which may develop ineffective strategies, especially when performance is suboptimal.

F   A Closer Look at Divergencies Between Gender Probing and Translation

In §4.2, we show that gender encoding correlates with the model's gender translation ability. For enc-dec, both gender encoding and gender translation achieve good accuracy, although the probe's accuracy is higher. As a result, some translations are incorrect even when gender can be accurately

predicted from the hidden states.

To better understand these discrepancies, we plot confusion matrices. Figure 3 shows the confusion matrix comparing *incorrect* probe predictions with the corresponding translations by the enc-dec model, while Figure 4 displays the confusion matrix for *correct* probe predictions alongside their associated translations. First, we observe that in Figure 3, the number of instances is low, as the probes are generally effective at predicting gender for enc-dec. Among the mismatches, the most frequent pattern involves the probe incorrectly predicting feminine while the model correctly translates masculine—23 cases for en→es, 22 for en→fr, and 8 for en→it. In contrast, Figure 4 shows that most discrepancies occur when the ST model translates as masculine despite the probe correctly predicting She: 33 cases for both en→es and en→fr, and 38 for en→it. The reverse case—feminine translations
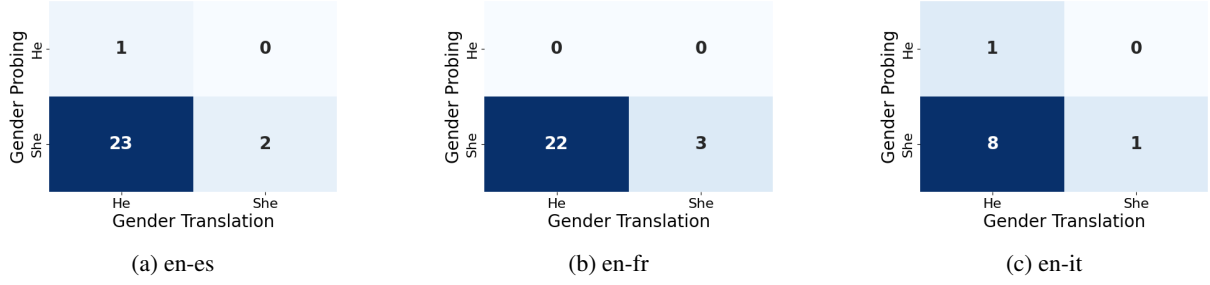
Figure 3: Confusion matrices showing the relationship between *incorrect* probe predictions (gender probing) and corresponding gender translation.
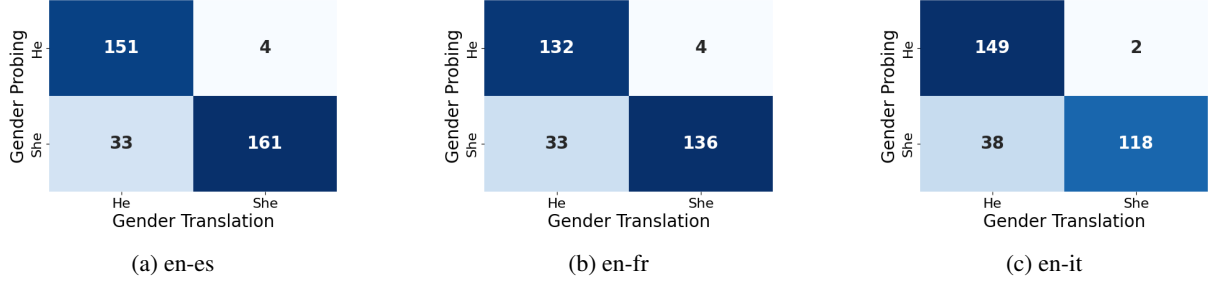

Figure 4: Confusion matrices showing the relationship between *correct* probe predictions (gender probing) and corresponding gender translation.

a.	en-es	She	SRC	They have formed me as a democratic citizen and a **bridge builder**.
			REF	Me han formado como ciudadana democrática y **constructora de puentes**.
			OUT	Me han formado como ciudadana democrática y como **un constructor de puente**.
b.	en-es	He	SRC	And just as the woman who wanted to know me as **an adult** got to know me, she turned into a [...].
			REF	Y así como la mujer que quería conocerme como **adulto** llegó a conocerme, se convirtió en una [...].
			OUT	Y así como una mujer que quería conocerme como **adulta** se conoció a mí, se convirtió en una [...].
c.	en-fr	He	SRC	And I wound up getting involved with the space community, really **involved** with NASA, [...].
			REF	Et j'ai fini par être impliqué dans la communauté spatiale, réellement **impliqué** avec la NASA, [...].
			OUT	Et j'ai fini par m'impliquer dans la communauté spatiale, vraiment **impliquée** par la NASA, [...].
d.	en-fr	She	SRC	My main sport was soccer, and I was **a goalkeeper**, which is [...].
			REF	Mon sport principal était le football et j'étais **gardienne de but**, ce qui est [...].
			OUT	Mon principal sport était le foot et j'étais **un gardien**, qui est [...].
e.	en-it	She	SRC	And I want to say something a little bit radical for **a feminist**, and that is that I think that [...].
			REF	Vorrei dire qualcosa di abbastanza radicale per **una femminista**, cioè che credo [...].
			OUT	E voglio dire qualcosa di un po' radicale per **un femminista**, e penso che [...].
f.	en-it	She	SRC	As a researcher, **a professor** and a new parent, my goal is to [...].
			REF	Da ricercatrice, **professoressa** e adesso genitore, il mio obiettivo è [...].
			OUT	Da ricercatrice, **professore** e genitore nuovo, il mio obiettivo è [...].

Table 6: Transcribed source sentences (SRC) from `test-speaker`, accompanied by reference translations (REF) and output translations (OUT) generated by `enc-dec`. Misgendered words in OUT, along with their correct counterparts in REF and aligned terms in SRC, are shown in bold.

for the He class—is much rarer, with only 4 cases each for en→es and en→fr, and 2 for en→it. This asymmetry reflects the persistent gender bias in translation, which affects even the enc-dec model despite its overall strong performance on gender translation.

To investigate why the ST model produces incorrect translations despite correctly encoding gender, we manually examined the divergent cases from Figure 4. We identified several patterns where the model struggles, and provide examples in Table 6. In certain cases, stereotypical associations—particularly involving professions—appear to influence the translations. For instance, in *a)*, *d)*, and *f)* terms like "bridge builder", "goalkeeper", and "professor"—often associated with men—are trans-

lated as masculine even when referring to women, despite other gendered words in the sentence being correctly translated as feminine. In other cases, the model appears to struggle with complex agreement resolutions. For example, in *c)*, "involved" may have been linked to "space community" due to proximity, rather than the speaker. In *e)*, "for a feminist" may not have been correctly interpreted as referring to the speaker. In *b)*, the presence of another gendered referent ("woman" and "she") may have led the model to misinterpret "adult" as feminine.

Overall, these findings suggest that linguistic biases can sometimes override acoustic cues, limiting the model's ability to fully leverage gender information. However, a more systematic analysis is needed in future work, which is beyond the scope of this paper.