FAMA: The First Large-Scale Open-Science Speech Foundation Model for English and Italian

Sara **Papi***, Marco **Gaido***, Luisa **Bentivogli**, Alessio **Brutti**, Mauro **Cettolo**, Roberto **Gretter**,
Marco **Matassoni**, Mohamed **Nabih** and Matteo **Negri**

Fondazione Bruno Kessler, Via Sommarive 18, 38123 Trento, Italy

Abstract

The development of speech foundation models (SFMs) like Whisper and SeamlessM4T has significantly advanced the field of speech processing. However, their closed nature–with inaccessible training data and code–poses major reproducibility and fair evaluation challenges. While other domains have made substantial progress toward open science by developing fully transparent models trained on open-source (OS) code and data, similar efforts in speech processing remain limited. To fill this gap, we introduce FAMA, the first family of open science SFMs for English and Italian, trained on 150k+ hours of OS speech data. Moreover, we present a new dataset containing 16k hours of cleaned and pseudo-labeled speech for both languages. Results show that FAMA achieves competitive performance compared to existing SFMs while being up to 8 times faster. All artifacts, including codebase, datasets, and models, are released under OS-compliant licenses, promoting openness in speech technology research. The FAMA collection is available at: https://huggingface.co/collections/FBK-MT/fama-683425df3fb2b3171e0cdc9e

1. Introduction

The development of speech foundation models (SFMs) has significantly advanced speech processing in the last few years, particularly in areas such as automatic speech recognition (ASR) and speech translation (ST). Popular SFMs such as OpenAI Whisper [1] and Meta SeamlessM4T [2] have been released to the public in various sizes and with extensive language coverage. However, these models completely lack comprehensive accessibility to their training codebases and datasets, hindering their reproducibility and raising concerns about potential data contamination [3], thereby complicating fair evaluation.

In other domains, multiple efforts towards building models that are more accessible, reproducible, and free from proprietary constraints have been made [4, 5, 6, 7, 8, 9, 10]. For instance, the OLMO project [11] has demonstrated the feasibility of training large language models (LLMs) using only open-source (OS) data [12], realizing an *open-science*[1] system [14] for text processing. However, such comprehensive approaches are still lacking in the field of speech processing.

Recent works towards this direction are represented by OWSM [15] and its subsequent versions [16]. OWSM, whose model weights and codebase used for the training are released open source, reproduces a Whisper-style training using publicly available data. Despite representing a valuable initiative toward building an open-science system, there is still a step missing for creating the first SFM of this kind: leveraging only data that is not only publicly available but also released under an OS-compliant license [17]. Such effort would allow users complete access and control over the data used at every stage of the scientific process, promoting reproducibility [18], fair evaluation [19], and the ability to build upon prior research without any barriers [20]. Besides transparency and collaboration, these efforts also foster users' trust by ensuring that data is not leveraged to build tools that can be used under conditions/purposes (e.g., commercial) for which the data was not intended [14].

To fill this gap, we release **FAMA**,[2] the first family of large-scale open-science SFMs for English and Italian trained on over 150k hours of exclusively OS-compliant

---

[1] *Open science* involves ensuring transparency and accessibility at all stages of the scientific process [13], including publishing OS research papers, data, code, and any information needed to replicate the research.

[2] *Fama* (from the Latin "fari" meaning "to speak") is the personification of the public voice in Roman mythology.

speech data. We leverage both already available OS datasets and create a new collection of ASR and ST psuedolabels for Italian and English comprising more than 16k hours of OS-compliant speech, along with automatically generated Italian and English translations for an additional 130k+ hours of speech. We also detail training and evaluation procedures and provide full access to training data to have complete control of the model creation and avoid data contamination issues. FAMA models achieve remarkable results, with up to 4.2 WER and 0.152 COMET improvement on average across languages compared to OWSM and remaining competitive in terms of ASR performance with the Whisper model family while being up to 8 times faster. All the artifacts used for realizing FAMA models, including codebase, datasets, and models themself, are released under OS-compliant licenses, promoting a more responsible creation of models in our community. Our approach would not only facilitate fair evaluation and comparison of SFMs but also encourage broader participation in speech technology development, leading to more inclusive and diverse applications.

The artifacts are available at:

🤗 **FAMA-medium (878M):**
https://hf.co/FBK-MT/fama-medium

🤗 **FAMA-small (479M):**
https://hf.co/FBK-MT/fama-small

🤗 **FAMA-medium-asr (878M):**
https://hf.co/FBK-MT/fama-medium-asr

🤗 **FAMA-small-asr (479M):**
https://hf.co/FBK-MT/fama-small-asr

🤗 **FAMA Training Data:**
https://hf.co/datasets/FBK-MT/fama-data

💠 **FAMA Code:**
https://github.com/hlt-mt/FBK-fairseq

2. The FAMA Framework

2.1. Training and Evaluation Data

In compliance with the open-science ideology, we train and test our models only on OS-compliant data. The training set comprises both already publicly available OS datasets, and new pseudolabels created for this work, whose list is presented in Table 1.

To create the new pseudolabels, we leveraged the speech content of YouTube-Commons,[3] a dataset collecting YouTube videos released with the permissive

³https://hf.co/datasets/PleIAs/YouTube-Commons

Dataset	#hours		Label
	en	*it*	
CommonVoice v18 [21]	1746	250	G
CoVoST2 [22]	420	28	G
FLEURS [23]	7	9	G
LibriSpeech [24]	358	-	G
MOSEL [17]	66,301	21,775	A
MLS [25]	44,600	247	G
VoxPopuli-ASR [26]	519	74	G
YouTube-Commons (*our paper*)	14,200	1,828	A
Total	**128,152**	**24,211**	G+A

Table 1
ASR: List of both publicly available training data and the data created in this paper for English (en) and Italian (it). "G" stands for gold labels while "A" for automatically generated labels (transcripts).

CC-BY 4.0 license. The videos are automatically converted into wav files with one channel and a sampling rate of 16k Hz. Then, the audio is cleaned from music and non-speech phenomena and segmented using silero [27], a lightweight VAD having low computational requirements. Lastly, to make it suitable for training, the audio is split using SHAS [28] in segments of around 16 seconds on average. The resulting dataset contains automatic transcripts, which we created with Whisper large-v3,[4] for 14,200 hours of speech for English (*en*) and 1,828 for Italian (*it*). Including publicly available data (113,951 hours for *en*, and 22,383 hours for *it*), the final ASR training set comprises 128,152 hours of *en* speech and 24,211 hours of *it* speech, with a total of 152,363 hours of speech data, including 48,259 gold-labeled hours.

Being composed of speech-transcript pairs, the data mentioned so far is suitable for ASR. For ST, instead, only CoVoST2 and FLEURS contain translations from and into *en* and *it*. For this reason, we automatically translated the transcripts of all the speech data (including the original CoVoST2) with MADLAD-400 3B-MT [29].[5] Following [30, 31], we additionally filter out samples based on the ratio r between the source and target text lengths (in characters) for each language pair based on their distribution ($r_{min} = 0.75$, $r_{max} = 1.45$ for en-it, and $r_{min} = 0.65$, $r_{max} = 1.35$ for it-en), resulting into 3.41% of data filtering for en-it and 3.12% for it-en. The final training set (Table 2) comprises the automatically translated speech data and the gold CoVoST2 and FLEURS datasets, resulting in a total of 147,686 hours for *en-it* and *it-en*.

For validation during training, and testing, we use gold-labeled benchmarks. ASR evaluation is conducted on CommonVoice, MLS, and VoxPopuli, with CommonVoice

⁴https://hf.co/openai/whisper-large-v3
⁵https://hf.co/google/madlad400-3b-mt

| Dataset | #hours | | Label |
	en-it	it-en	
CommonVoice v18 [21]	1746	250	A
CoVoST2 [22] - automatic labels	420	28	A
LibriSpeech [24]	358	-	A
MOSEL [17]	66,301	21,775	A
MLS [25]	44,600	247	A
VoxPopuli-ASR [26]	519	74	A
YouTube-Commons (*our paper*)	14,200	1,828	A
Total (A)	128,144	24,202	A
Filtered (A)	123,777	23,445	A
CoVoST2 [22] - gold labels	420	28	G
FLEURS [23]	7	9	G
Total	**124,204**	**23,482**	G+A

Table 2

ST: List of both publicly available training data and the data created in this paper for English-Italian (en-it) and Italian-English (it-en). "G" stands for gold labels while "A" for automatically generated labels (translations).

also serving as the validation set for both *en* and *it*. For translation, we use CoVoST2 for *it-en* and FLEURS dev and test sets for *en-it*.

2.2. Model Architecture

FAMA models are two-sized encoder-decoder architectures, `small` and `medium`. Both models are composed of a Conformer encoder [32] and a Transformer decoder [33]. FAMA `small` has 12 encoder layers and 6 decoder layers, while FAMA `medium` has 24 encoder layers and 12 decoder layers. Our decision to use an encoder twice as deep as the decoder–unlike Whisper and OWSM, which have an equal number of encoder and decoder layers–is driven by two key motivations: *i)* since autoregressive models perform multiple decoder passes during output generation, a shallower decoder speeds up inference by making each pass faster, and *ii)* since many approaches integrate SFMs with LLMs by leveraging the encoder [34], a deeper encoder helps preserve more of the SFMs processing capabilities in such integrations. Each layer has 16 attention heads, an embedding dimension of 1,024, and an FFN dimension of 4,096.

The Conformer encoder is preceded by two 1D convolutional layers with a stride of 2 and a kernel size of 5. The kernel size of the Conformer convolutional module is 31 for both the point- and depth-wise convolutions. The vocabulary is built using a SentencePiece unigram model [35] with size 16,000 trained on *en* and *it* transcripts. Two extra tokens–`<lang:en>` and `<lang:it>`–are added to indicate whether the target text is in *en* or *it*. The input audio is represented by 80 Mel-filterbank features extracted every 10 ms with a window of 25 ms.

2.3. Training and Evaluation Procedures

We train both models using a combination of three losses. First, a label-smoothed cross-entropy loss ($\mathcal{L}_{CE}$) is applied to the decoder output, using the target text as the reference (transcripts for ASR and translations for ST). Second, a CTC loss [36] is computed using transcripts as reference ($\mathcal{L}_{CTCsrc}$) on the output of the 8th encoder layer for `small` and the 16th for `medium`. Third, a CTC loss on the final encoder output ($\mathcal{L}_{CTCtgt}$) is applied to predict the target text. The final loss is the weighted sum of the above-mentioned losses:

$$\mathcal{L} = \lambda_1 \mathcal{L}_{CE} + \lambda_2 \mathcal{L}_{CTCsrc} + \lambda_3 \mathcal{L}_{CTCtgt}$$

where $\lambda_1, \lambda_2, \lambda_3 = 5.0, 1.0, 2.0$, and the label smoothing factor of the CE is 0.1.

FAMA models are trained using a two-stage approach, where the model is pre-trained first on ASR data only (ASR pre-training) and then trained on both ASR and ST data (ASR+ST training). Both training stages lasted 1M steps, corresponding to ~6 epochs over the training data.

For the ASR pre-training, the learning rate (lr_{S1}) scheduler adopted to train the `small` model is the Noam scheduler [33] with a peak of 2e-3 and 25,000 warm-up steps. To cope with convergence issues similar to [16], for the `medium` model we adopted a piece-wise warm-up on the Noam scheduler, with the learning rate first increasing linearly to 2e-5 for 25k steps and then to 2e-4 for an additional 25k steps, followed by the standard inverse square root function. For the ASR+ST training, we sample the ASR target with probability p_{ASR}=0.5 and use the ST target otherwise. Training settings are the same as for ASR pre-training, except for the learning rate that is set to a constant value lr_{S2}=1e-4. Experiments on how p_{ASR} and lr_{S2} are determined for the `small` model are discussed in Section 3.1. For the `medium` model, similarly to the first stage, the lr_{S2} is scaled down by one order of magnitude compared to the `small` model, i.e., a constant value lr_{S2}=1e-5 is used.

The optimizer is AdamW with momentum $\beta_1, \beta_2 = 0.9, 0.98$, a weight decay of 0.001, a dropout of 0.1, and clip normalization of 10.0. We apply SpecAugment [37] during both ASR pre-training and ASR+ST training. We use mini-batches of 10,000 tokens for FAMA `small` and 4,500 for FAMA `medium` with an update frequency of, respectively, 2 and 6 on 16 NVIDIA A100 GPUs (64GB RAM), save checkpoints every 1,000 steps and average the last 25 checkpoints to obtain the final model.

The inference is performed using a single NVIDIA A100 GPU with a batch size of 80,000 tokens. We use beam search with beam 5, unknown penalty of 10,000, and no-repeat n-gram size of 5. Additionally, we report the results using the joint CTC rescoring [38], leveraging the CTC on the encoder output with weight 0.2. Both training and inference are done using the bug-free Con-

former implementation [39] available in FBK-fairseq,[6] which is built upon fairseq-S2T [40]. ASR performance is evaluated with word error rate (WER) using the jiWER library[7] with the text normalized using Whisper normalizer[8]. ST performance is evaluated using COMET [41] version 2.2.4, with the default `Unbabel/wmt22-comet-da` model.

2.4. Terms of Comparison

As a first term of comparison, we use Whisper [1] in both `medium`[9] and `large-v3` configurations as the first is comparable with FAMA `medium` in terms of size and the second–trained on more than 4M hours—is the best performing model of the Whisper family. The comparison is made for *en* and *it* ASR and *it-en* ST, as Whisper does not cover the *en-to-many* translation directions. Whisper models are released under Apache 2.0 license and, therefore, open weights. For both ASR and ST, we also compare with SeamlessM4T `medium`[10] and `v2-large`[11] covering ASR and both ST language directions [2]. The model is non-commercial and, therefore, not open. We also compare with OWSM `v3.1 medium`[12], the best performing model of the OWSM family, also covering ASR and both ST language directions and released open source [16].
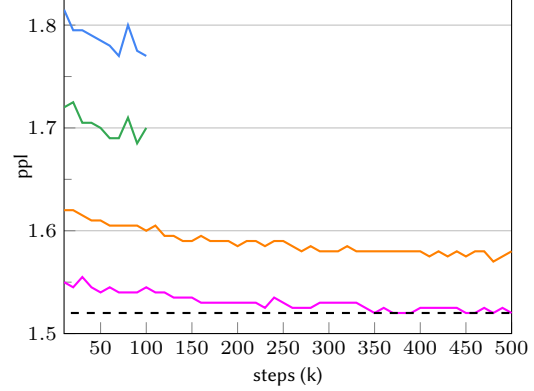
To ensure a fair comparison, we perform the inference with HuggingFace transformers[13] version 4.48.1 using the standard settings and beam search with beam 5, except for OWSM, which is not supported on HuggingFace, and for which the original ESPNet[14] inference code is used with a beam size of 3.[15]
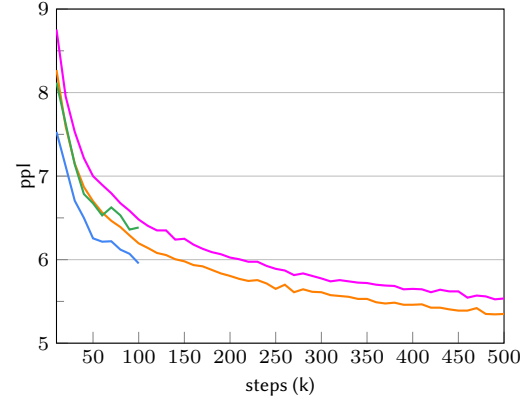
3. Results

3.1. Pre-training and Catastrophic Forgetting

Catastrophic forgetting is a well-known problem in machine learning [42] that arises when a system is trained sequentially on multiple languages or tasks, leading to a degradation in performance on original domains or languages [43]. As we follow a two-stage approach, which is commonly employed in SFMs training [1], we analyze

---

[6]https://github.com/hlt-mt/FBK-fairseq

[7]https://pypi.org/project/jiwer/

[8]https://pypi.org/project/whisper-normalizer/

[9]https://hf.co/openai/whisper-medium

[10]https://hf.co/facebook/hf-seamless-m4t-medium

[11]https://hf.co/facebook/seamless-m4t-v2-large

[12]https://hf.co/espnet/owsm_v3.1_ebf

[13]https://pypi.org/project/transformers/

[14]https://github.com/espnet/espnet/tree/master/egs2/owsm_v3.1/s2t1

[15]We attempted to use a beam size of 5 but the model had out-of-memory issues even when reducing the batch size.

---

the conditions in which this phenomenon arises during the ASR+ST training.

(a) ASR

(b) ST

lr_{S2}=1e-3, p_{ASR}=0.2 lr_{S2}=1e-4, p_{ASR}=0.2
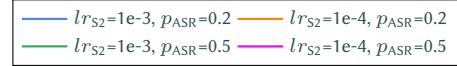lr_{S2}=1e-3, p_{ASR}=0.5 lr_{S2}=1e-4, p_{ASR}=0.5

Figure 1: Average ASR and ST perplexity (ppl) on both English and Italian up to 500k steps of the training. Due to the evident worse results achieved by using a lr of 1e-3, we stopped the training curves after 100k steps. The black dashed line is the ppl of the ASR model from which the training is started.

Figure 1 shows the perplexity (ppl) behavior during the first 100/500k steps of the FAMA `small` model training on the validation sets. We present the results of different systems obtained by varying both the learning rate lr_{S2} and the sampling probability p_{ASR} discussed in Section 2.3. Lower values of lr_{S2} (e.g., 1e-5) lead to worse performance and are not included in the results. Since the computational budget for our experiments is limited, we analyze two cases for the sampling probability: 1) p_{ASR}=0.5 to obtain a system equally trained on both ASR and ST tasks, and 2) p_{ASR}=0.2 to obtain a system trained

Model	#params	ASR (WER ↓)								ST (COMET ↑)	
		CV		MLS		VP		AVG		CVST2	FLRS
		en	it	en	it	en	it	en	it	it-en	en-it
Whisper medium	769M	14.5	10.4	14.2	15.9	8.1	26.8	12.3	17.7	0.801	-
Whisper large-v3	1550M	11.2	6.5	**5.0**	8.8	7.1	18.8	7.8	11.4	0.825	-
OWSM v3.1 medium	1020M	11.9	12.5	6.6	19.3	8.4	24.0	9.0	18.6	0.636	0.337
SeamlessM4T medium	1200M	10.7	7.8	8.8	11.3	10.2	18.2	9.9	12.4	0.831	0.820
SeamlessM4T v2-large	2300M	**7.7**	**5.0**	6.4	**8.5**	**6.9**	16.6	**7.0**	**10.0**	**0.852**	**0.855**
FAMA-ASR small	475M	13.8	8.9	5.8	12.6	7.2	15.7	8.9	12.4	-	-
+ joint CTC rescoring		13.9	8.9	5.8	12.4	7.0	14.6	8.9	12.0	-	-
FAMA-ASR medium	878M	11.7	7.1	5.1	12.2	7.0	15.9	7.9	11.7	-	-
+ joint CTC rescoring		11.7	7.0	5.1	12.2	7.0	14.6	7.9	11.3	-	-
FAMA small	475M	13.7	8.6	5.8	12.8	7.3	15.6	8.9	12.3	0.774	0.807
+ joint CTC rescoring		13.6	8.5	5.8	12.8	7.2	14.8	8.9	12.0	0.777	0.804
FAMA medium	878M	11.5	7.0	5.2	13.9	7.2	15.9	8.0	12.3	0.787	0.821
+ joint CTC rescoring		11.5	7.7	5.2	13.5	7.1	14.9	7.9	12.0	0.791	0.818

Table 3
ASR and ST performance of FAMA models and existing SFMs as terms of comparison. The results are reported on CommonVoice (CV), Multilingual LibriSpeech (MLS), and VoxPopuli (VP) for ASR, and on CoVoST (CVST2), and FLEURS (FLRS) for ST. Best values are in bold.

more on the unseen task during pre-training, i.e., the ST task.

As we can see from the curves, a lr_{S2} of 1e-3 seems to be too high for maintaining good ASR performance while learning a new task (ST). Both in the case in which the ST training is more boosted (p_{ASR}=0.2) and in the case in which ASR and ST training is balanced (p_{ASR}=0.5), we notice a significant increase in the ASR ppl of up to 0.25 that corresponds to a drop in performance of 3-4 WER on both languages – which, moreover, is not recovered later on in the training. Therefore, to avoid catastrophic forgetting arising just in the first steps, we exclude lr_{S2}=1e-3 and use 1e-4 for the two-stage training. Regarding the ASR sampling, we look at the behavior of the curves for 500k steps (half of the second-stage training) and notice that the ASR ppl curve with p_{ASR}=0.5 slowly approaches the original model ppl value while the one with p_{ASR}=0.2, despite improving, is not able to approach the original ppl value. This is counterbalanced by a lower (hence, better) ppl of the p_{ASR}=0.2 curve on ST compared to that of the p_{ASR}=0.5 curve. However, this difference, which is about ∼0.2 ppl, is not reflected in the ST performance, which only improves by 0.005 COMET points on average. Instead, the difference in terms of WER is significant, with a quality drop of ∼0.8 WER across en and it. As a result, we conclude that we avoid catastrophic forgetting in the two-stage training only by evenly sampling the ASR and ST tasks during the second step.

3.2. Comparison with Existing SFMs

In Table 3, we show the results for both ASR and ST of our FAMA models and SFMs presented in Section 2.4. For FAMA models, we provide the scores of the ASR-only model (FAMA-ASR), obtained after pre-training, and of the final ASR+ST model, as well as the results obtained through joint CTC rescoring.

Looking at the results of FAMA-ASR, we observe that the medium model outperforms the small one, with ∼0.8 WER improvements on average both with and without the joint CTC rescoring. Compared to Whisper medium, FAMA achieves better results with FAMA medium outperforming Whisper by 4.4 WER on *en* and 6.4 on *it* while having a similar number of model parameters. Remarkable performance is achieved by FAMA medium also compared to OWSM v3.1 medium, with improvements of up to 1.1 WER on *en* and 7.3 on *it*, but also compared to Whisper large-v3, where similar WER scores are achieved. Instead, SeamlessM4T models, leveraging large pretrained models such as wav2vec-BERT 2.0 (which is trained on 4.5 million hours) and NLLB (which is trained on more than 43 billion sentences), still outperform FAMA, with the v2-large scoring an incredibly low WER on CommonVoice also compared to a strong competitor as Whisper large-v3. Looking at the ASR results of the final FAMA models, we observe that the WER remained almost unaltered compared to the ASR-only model, as already discussed in Section 3.1. Regarding ST results, we notice that FAMA models outperform OWSM v3.1 medium, with an improvement of up to 0.141 COMET by FAMA small and 0.152 by FAMA medium while still struggling to achieve the performance of Whisper and SeamlessM4T.

These mixed outcomes–competitive ASR performance even against larger non-open models but lower ST performance–demonstrate both the feasibility of building high-quality open-science SFMs and the need for initiatives dedicated to creating OS-compliant ST datasets with human references to bridge the gap with non-open

Model	Batch Size	xRTF (↑)		
		en	*it*	*AVG*
Whisper `medium`	8	13.3	10.9	12.1
Whisper `large-v3`	4	7.9	6.5	7.2
SeamlessM4T `medium`	2	28.5	26.2	27.4
SeamlessM4T `v2-large`	2	13.7	13.3	13.5
FAMA `small`	16	**57.4**	**56.0**	**56.7**
FAMA `medium`	8	39.5	41.2	40.4

Table 4

Computational time and maximum batch size for Whisper, SeamlessM4T, and FAMA models. Best values are in bold.

models.

3.3. Computational Time

As an additional comparison, we evaluate the throughput of the SFMs on a single NVIDIA A40 40GB. The throughput, measured in xRTF (the inverse of the real-time factor),[16] is calculated as the number of seconds of processed audio divided by the compute time in seconds. The test set used for this performance evaluation is CommonVoice on both *en* and *it* with a total duration of, respectively, 26.9 and 26.4 hours. For each model, we report the maximum batch size possible spanning in the range 2, 4, 8, and 16, as higher values resulted in out-of-memory issues with all models. The results are reported in Table 4.

We notice that Whisper models are the slowest ones, with an average xRTF of 12.1 for `medium` and 7.2 for `large-v3`, making them ∼3-6 times slower than FAMA `medium` and ∼5-8 than FAMA `small`. These results can be attributed to the architectural design of Whisper models that apply an ×2 audio subsampling compared to the commonly used ×4 (as in FAMA) and introduce a lot of padding in shorter sequences to achieve the fixed 30-second length. The Seamless models, despite having no extra padding (as FAMA) and a greater audio subsampling of ×8, are ∼2 times faster than Whisper ones but still 1.5-3 times slower for, respectively, `medium` and `v2-large`, compared to FAMA `medium` and 2-4 compared to FAMA `small`, making the FAMA model family the fastest by a large margin.

3.4. Gender Bias Analysis

We also measure the gender bias disparity between male and female performance using the ASR benchmark proposed by Attanasio et al. [44]. The results are presented in Table 5[17] and are measured as absolute performance gaps

---

Model	Gap R	Gap S	AVG
Whisper `large-v3`	0.5584	**0.9711**	**0.7648**
SeamlessM4T `v2-large`	0.4485	2.3271	1.3878
FAMA-ASR `small`	**0.0250**	1.7191	0.8721
FAMA-ASR `medium`	0.4074	2.0558	1.2316
FAMA `small`	0.7569	1.5642	1.1605
FAMA `medium`	0.2165	1.7661	0.9913

Table 5

Absolute WER quality gaps between female and male subsets, divided into read (Gap R) and spontaneous (Gap S) speech.

between female WER and male WER scores obtained on CommonVoice 17 and VoxPopuli.

We can observe that FAMA-ASR `small` obtained the smallest–hence, best–performance gap between male and feminine transcription from read speech, with a gap being an order of magnitude smaller than all other models. When moving to the spontaneous speech, instead, Whisper `large-v3` obtains the best result. Overall, Whisper achieves the best average result, followed by FAMA-ASR `small` and FAMA `medium`, which are the only models scoring less than a 1.0 WER difference. All FAMA models can outperform Seamless M4T `v2-large`, achieving an average gap reduction of 0.16 to 0.52.

4. Conclusions

In this paper, we addressed the challenges posed by the closed nature of existing SFMs, such as limited accessibility to training data and codebases, by introducing FAMA, the first large-scale open-science SFM for English and Italian. Trained on over 150k hours of exclusively OS speech, FAMA ensures full transparency, with all artifacts released under OS-compliant licenses. Additionally, we contributed a new collection of ASR and ST pseudolabels for about 16k hours of speech data, and more than 130k hours of English and Italian automatic translations. Results show that FAMA models outperform OWSM on both ASR and ST and also achieve comparable ASR results to Whisper while being up to 8 times faster. By providing the community with fully accessible

resources, FAMA bridges the gap between advances in speech technology and open science principles, enabling fair evaluation, broader participation, and inclusivity. Future work will focus on extending FAMA to additional languages with the ultimate goal of further expanding the open science ecosystem to speech technologies.

Acknowledgments

References

[1] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International Conference on Machine Learning, PMLR, 2023, pp. 28492–28518.

[2] L. Barrault, Y.-A. Chung, M. C. Meglioli, D. Dale, N. Dong, P.-A. Duquenne, H. Elsahar, H. Gong, K. Heffernan, J. Hoffman, et al., Seamlessm4t: Massively multilingual & multimodal machine translation, arXiv preprint arXiv:2308.11596 (2023).

[3] Y. Dong, X. Jiang, H. Liu, Z. Jin, B. Gu, M. Yang, G. Li, Generalization or memorization: Data contamination and trustworthy evaluation for large language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Findings of the Association for Computational Linguistics: ACL 2024, Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 12039–12050.

[4] BigScience Workshop, T. L. Scao, A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, et al., Bloom: A 176b-parameter open-access multilingual language model, arXiv preprint arXiv:2211.05100 (2022).

[5] S. Biderman, H. Schoelkopf, Q. Anthony, H. Bradley, K. O'Brien, E. Hallahan, M. A. Khan, S. Purohit, U. S. Prashanth, E. Raff, A. Skowron, L. Sutawika, O. Van Der Wal, Pythia: a suite for analyzing large language models across training and scaling, in: Proceedings of the 40th International Conference on Machine Learning, ICML'23, JMLR.org, 2023.

[6] Z. Liu, A. Qiao, W. Neiswanger, H. Wang, B. Tan, T. Tao, J. Li, Y. Wang, S. Sun, O. Pangarkar, et al., Llm360: Towards fully transparent open-source llms, in: First Conference on Language Modeling, 2024.

[7] Q. Sun, Y. Luo, S. Li, W. Zhang, W. Liu, OpenOmni: A collaborative open source tool for building future-ready multimodal conversational agents, in: D. I. Hernandez Farias, T. Hope, M. Li (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 46–52.

[8] M. Deitke, C. Clark, S. Lee, R. Tripathi, Y. Yang, J. S. Park, M. Salehi, N. Muennighoff, K. Lo, L. Soldaini, et al., Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models, arXiv preprint arXiv:2409.17146 (2024).

[9] W. Dai, N. Lee, B. Wang, Z. Yang, Z. Liu, J. Barker, T. Rintamaki, M. Shoeybi, B. Catanzaro, W. Ping, Nvlm: Open frontier-class multimodal llms, arXiv preprint arXiv:2409.11402 (2024).

[10] P. H. Martins, P. Fernandes, J. Alves, N. M. Guerreiro, R. Rei, D. M. Alves, J. Pombal, A. Farajian, M. Faysse, M. Klimaszewski, P. Colombo, B. Haddow, J. G. de Souza, A. Birch, A. F. Martins, Eurollm: Multilingual language models for europe, Procedia Computer Science 255 (2025) 53–62. Proceedings of the Second EuroHPC user day.

[11] D. Groeneveld, et al., OLMo: Accelerating the science of language models, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 15789–15809.

[12] L. Soldaini, et al., Dolma: an open corpus of three trillion tokens for language model pretraining research, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 15725–15788.

[13] R. Vicente-Saez, C. Martinez-Fuentes, Open science now: A systematic literature review for an integrated definition, Journal of Business Research 88 (2018) 428–436.

[14] M. White, I. Haddad, C. Osborne, X.-Y. Y. Liu, A. Abdelmonsef, S. Varghese, A. L. Hors, The model openness framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence, arXiv preprint arXiv:2403.13784 (2024).

[15] Y. Peng, J. Tian, B. Yan, D. Berrebbi, X. Chang, X. Li, J. Shi, S. Arora, W. Chen, R. Sharma, W. Zhang, Y. Sudo, M. Shakeel, J.-W. Jung, S. Maiti, S. Watanabe, Reproducing whisper-style training using an open-source toolkit and publicly available data, in:

2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), 2023, pp. 1–8.

[16] Y. Peng, J. Tian, W. Chen, S. Arora, B. Yan, Y. Sudo, M. Shakeel, K. Choi, J. Shi, X. Chang, J. weon Jung, S. Watanabe, Owsm v3.1: Better and faster open whisper-style speech models based on e-branchformer, in: Interspeech 2024, 2024, pp. 352–356.

[17] M. Gaido, S. Papi, L. Bentivogli, A. Brutti, M. Cettolo, R. Gretter, M. Matassoni, M. Nabih, M. Negri, MOSEL: 950,000 hours of speech data for open-source speech foundation model training on EU languages, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 13934–13947.

[18] A. Belz, C. Thomson, E. Reiter, S. Mille, Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP, in: A. Rogers, J. Boyd-Graber, N. Okazaki (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 3676–3687.

[19] S. Balloccu, P. Schmidtová, M. Lango, O. Dusek, Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs, in: Y. Graham, M. Purver (Eds.), Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, St. Julian's, Malta, 2024, pp. 67–93.

[20] H. Chesbrough, From open science to open innovation, Institute for Innovation and Knowledge Management, ESADE (2015).

[21] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, G. Weber, Common voice: A massively-multilingual speech corpus, in: Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), 2020, pp. 4211–4215.

[22] C. Wang, A. Wu, J. Gu, J. Pino, CoVoST 2 and Massively Multilingual Speech Translation, in: Proc. Interspeech 2021, 2021, pp. 2247–2251.

[23] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, A. Bapna, Fleurs: Few-shot learning evaluation of universal representations of speech, in: 2022 IEEE Spoken Language Technology Workshop (SLT), 2023, pp. 798–805.

[24] V. Panayotov, G. Chen, D. Povey, S. Khudanpur, Librispeech: An ASR corpus based on public domain audio books, in: 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 5206–5210.

[25] V. Pratap, Q. Xu, A. Sriram, G. Synnaeve, R. Collobert, MLS: A Large-Scale Multilingual Dataset for Speech Research, in: Proc. Interspeech 2020, 2020, pp. 2757–2761.

[26] C. Wang, M. Riviere, A. Lee, A. Wu, C. Talnikar, D. Haziza, M. Williamson, J. Pino, E. Dupoux, VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation, in: C. Zong, F. Xia, W. Li, R. Navigli (Eds.), 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), Association for Computational Linguistics, Online, 2021, pp. 993–1003.

[27] S. Team, Silero vad: pre-trained enterprise-grade voice activity detector (vad), number detector and language classifier, https://github.com/snakers4/silero-vad, 2024.

[28] I. Tsiamas, G. I. Gállego, J. A. R. Fonollosa, M. R. Costa-jussà, Shas: Approaching optimal segmentation for end-to-end speech translation, in: Interspeech 2022, 2022, pp. 106–110.

[29] S. Kudugunta, I. Caswell, B. Zhang, X. Garcia, D. Xin, A. Kusupati, R. Stella, A. Bapna, O. Firat, Madlad-400: a multilingual and document-level large audited dataset, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[30] M. Gaido, S. Papi, D. Fucci, G. Fiameni, M. Negri, M. Turchi, Efficient yet competitive speech translation: FBK@IWSLT2022, in: E. Salesky, M. Federico, M. Costa-jussà (Eds.), Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), Association for Computational Linguistics, Dublin, Ireland (in-person and online), 2022, pp. 177–189.

[31] M. M. I. Alam, A. Anastasopoulos, A case study on filtering for end-to-end speech translation, arXiv preprint arXiv:2402.01945 (2024).

[32] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, R. Pang, Conformer: Convolution-augmented Transformer for Speech Recognition, in: Proc. Interspeech 2020, 2020, pp. 5036–5040.

[33] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. u. Kaiser, I. Polosukhin, Attention is all you need, in: I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 30, 2017. URL: https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[34] M. Gaido, S. Papi, M. Negri, L. Bentivogli, Speech

translation with speech foundation models and large language models: What is there and what is missing?, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 14760–14778.

[35] T. Kudo, J. Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing, in: E. Blanco, W. Lu (Eds.), Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 66–71.

[36] A. Graves, S. Fernández, F. Gomez, J. Schmidhuber, Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks, in: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, New York, NY, USA, 2006, p. 369–376.

[37] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, Q. V. Le, SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition, in: Proc. Interspeech 2019, 2019, pp. 2613–2617.

[38] B. Yan, S. Dalmia, Y. Higuchi, G. Neubig, F. Metze, A. W. Black, S. Watanabe, CTC alignments improve autoregressive translation, in: A. Vlachos, I. Augenstein (Eds.), Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1623–1639.

[39] S. Papi, M. Gaido, A. Pilzer, M. Negri, When good and reproducible results are a giant with feet of clay: The importance of software quality in NLP, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 3657–3672.

[40] C. Wang, Y. Tang, X. Ma, A. Wu, D. Okhonko, J. Pino, 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, 2020.

[41] R. Rei, C. Stewart, A. C. Farinha, A. Lavie, COMET: A neural framework for MT evaluation, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online, 2020, pp. 2685–2702.

[42] M. McCloskey, N. J. Cohen, Catastrophic interference in connectionist networks: The sequential learning problem, volume 24 of *Psychology of Learning and Motivation*, 1989.

[43] S. Kar, G. Castellucci, S. Filice, S. Malmasi, O. Rokhlenko, Preventing catastrophic forgetting in continual learning of new natural language tasks, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 3137–3145.

[44] G. Attanasio, B. Savoldi, D. Fucci, D. Hovy, Twists, humps, and pebbles: Multilingual speech recognition models exhibit gender performance gaps, in: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 21318–21340.