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

Anonymous ACL submission

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 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 code, datasets, and models, will be released under OS-compliant licenses, promoting openness in speech technology research.

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 (Radford et al., 2023) and Meta SeamlessM4T (Barrault et al., 2023) 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 (Dong et al., 2024), 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 (BigScience Workshop et al., 2022; Biderman et al., 2023; Liu et al., 2023; Sun et al., 2024; Deitke et al., 2024; Dai et al., 2024; Martins et al., 2024). For instance, the OLMO project (Groeneveld et al., 2024) has demonstrated the feasibility of training large language models (LLMs) using only open-source (OS) data (Soldaini et al., 2024), realizing an *open-science*[1] system (White et al., 2024) for text processing. However, such comprehensive approaches are still lacking in the field of speech processing.

Recent works towards this direction include OWSM (Peng et al., 2023) and its subsequent versions (Peng et al., 2024). OWSM, whose model weights and the 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 (Gaido et al., 2024a). Such effort would allow users complete access and control over the data used at every stage of the scientific process, promoting reproducibility (Belz et al., 2023), fair evaluation (Balloccu et al., 2024), and the ability to build upon prior research without any barriers (Chesbrough, 2015). 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 (White et al., 2024).

To fill this gap, we create **FAMA**,[2] the first family of large-scale open-science SFMs for English

[1] *Open science* involves ensuring transparency and accessibility at all stages of the scientific process (Vicente-Saez and Martinez-Fuentes, 2018), including publishing OS research papers, data, and code needed to replicate the research.

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

and Italian trained on over 150k hours of exclusively OS-compliant 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 will be released, upon paper acceptance, 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.

2 The FAMA Framework

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 public OS datasets (Gaido et al., 2024a), and new pseudolabels, detailed in Appendix A. The resulting ASR train set comprises 152k hours, 128k for English (*en*) and 24k for Italian (*it*), and 174k hours for ST, 150k for *en-it*, and 24k for *it-en*. To validate and test, we use gold-labeled data. ASR evaluation is conducted on CommonVoice, MLS, and VoxPopuli, with CommonVoice 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*.

Model Architecture. FAMA models are two-sized encoder-decoder architectures, `small` and `medium`. Both models are composed of a Conformer encoder (Gulati et al., 2020) and a Transformer decoder (Vaswani et al., 2017). 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 (Gaido et al., 2024b), 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 a FFN dimension of 4,096. Full architectural specifications are provided in Appendix B.

Model Training and Evaluation. We adopt a two-stage training pipeline: ASR pre-training followed by multitask ASR+ST fine-tuning. For ASR pre-training, we use the Noam learning rate scheduler (Vaswani et al., 2017). To address convergence issues observed in larger speech models (Peng et al., 2024), we apply a modified warm-up strategy for FAMA `medium`, as detailed in Appendix C. During ASR+ST training, the model alternates between ASR and ST targets, sampling ASR with probability p_{ASR}=0.5. This value, along with the learning rate lr_{S2}, is selected based on validation perplexity trends discussed in §3. To reduce instability in larger models, we scale down the learning rate for FAMA `medium` by an order of magnitude. ASR performance is evaluated with word error rate (WER) using the jiWER library[3] with the text normalized using Whisper normalizer[4]. ST performance is evaluated using COMET (Rei et al., 2020) version 2.2.4, with the default `Unbabel/wmt22-comet-da` model. Full hyperparameters, training, and evaluation procedures are presented in Appendix C.

Terms of Comparison. As a first term of comparison, we use Whisper (Radford et al., 2023) `medium` and `large-v3` as the former is comparable to FAMA `medium` in size, and the latter–trained on over 4M hours–is the best model in the Whisper family. They are released under the Apache 2.0 license and thus have open weights. For both ASR and ST, we also compare with SeamlessM4T `medium` and `v2-large` covering ASR and both ST language directions (Barrault et al., 2023). The model is non-commercial, thus not open. We also compare with OWSM `v3.1 medium`, the best model in the OWSM family, covering both ASR

[3] https://pypi.org/project/jiwer/
[4] https://pypi.org/project/whisper-normalizer/

and ST language directions and released open source (Peng et al., 2024).

3 Results

Pre-training and Catastrophic Forgetting. Catastrophic forgetting occurs when sequential training on multiple tasks or languages leads to performance degradation on earlier ones (McCloskey and Cohen, 1989; Kar et al., 2022). Since we follow a two-stage training setup, common in SFM development (Barrault et al., 2023), we analyze when forgetting arises during ASR+ST training. Figure 1 shows perplexity (ppl) curves over the first 100/500k steps of FAMA `small` under different learning rates (lr_{S2})[5] and ASR sampling probabilities (p_{ASR}) (§2). Due to limited compute, we consider two sampling settings: p_{ASR}=0.5 for equal ASR/ST training, and p_{ASR}=0.2 to emphasize the unseen ST task. From the curves, we observe that a learning rate (lr_{S2}) of 1e-3 is too high to maintain good ASR performance while learning the new ST task. In both settings–boosting ST training (p_{ASR}=0.2) and balancing ASR and ST training (p_{ASR}=0.5)–we see a notable increase in ASR perplexity (up to +0.25), which translates to a 3-4 WER drop across both languages. Crucially, this degradation is not recovered later in training. To avoid catastrophic forgetting in the early stages, we discard lr_{S2}=1e-3 and adopt 1e-4 for the second stage. Looking at ASR sampling, we analyze the ppl curves after 500k steps (halfway through the second stage). With p_{ASR}=0.5, the ASR ppl slowly converges back toward the original value, while with p_{ASR}=0.2, it remains higher despite some improvement. Although p_{ASR}=0.2 yields slightly better ST perplexity (by ∼0.2), this does not translate into meaningful gains in downstream performance–only a marginal +0.005 COMET improvement on average. Meanwhile, the ASR degradation is more substantial, with ∼0.8 WER loss across both languages. We conclude that avoiding catastrophic forgetting during two-stage training requires evenly sampling ASR and ST targets in the second step.

Comparison with Existing SFMs. In Table 1, we show the results for both ASR and ST of our FAMA models and SFMs presented in §2. For FAMA models, we provide the scores of the ASR-only model (FAMA-ASR), and of the final

[5]Lower values of lr_{S2} (e.g., 1e-5) underperform and are excluded.

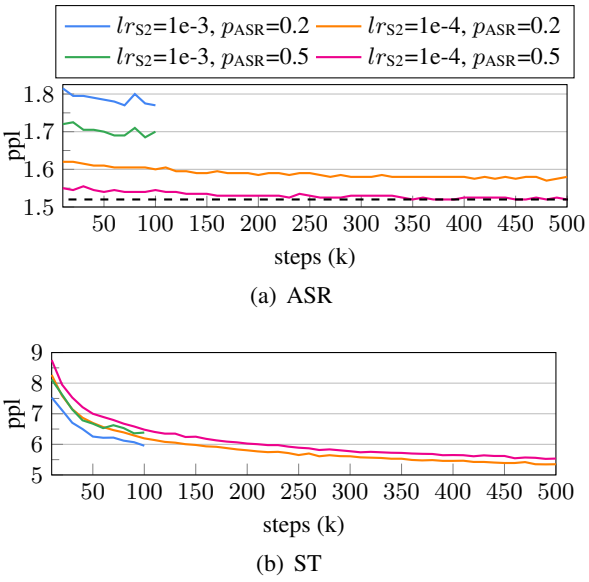

Figure 1: Average ASR and ST ppl up to 500k steps of the training. Due to the evident worse results achieved by using a lr of 1e-3, we stopped the training curves after 100k steps. The black dashed line is the ppl of the ASR model from which the training is started.

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 discussed before. 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

Model	#params	ASR (WER ↓)								ST (COMET ↑)	
		CommonVoice		MLS		VoxPopuli		*AVG*		CoVoST2	FLEURS
		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 1: ASR and ST results of FAMA models and existing SFMs as terms of comparison. Best values are **bold**.

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 models.

Computational Time. As an additional comparison, we evaluate the throughput of the SFMs, measured in xRTF (the inverse of the real-time factor),[6] which is calculated as the number of seconds of processed audio divided by the compute time in seconds. For each model,[7] 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 2. 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.

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 2: Computational time and maximum batch size on CommonVoice *en* and *it*. Best values are **bold**.

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.

[6] https://github.com/NVIDIA/DeepLearningExamples/blob/master/Kaldi/SpeechRecognition/README.md

[7] The benchmarking is performed on HuggingFace, thus we excluded OWSM from the comparison as it is not supported.

Limitations

Our work focuses on two European languages (English and Italian) and a limited set of translation directions, which may not generalize to typologically distant or low-resource language pairs. Due to computational constraints, we evaluate only two model sizes (`small` and `medium`) and explore a narrow range of hyperparameters. The training data are also limited compared to large-scale training on non-open data (Radford et al., 2023; Barrault et al., 2023), which might affect the broader generality of our findings. Finally, we do not investigate integration with large language models, which is a promising direction for future work (Gaido et al., 2024b).

Potential Risks. Although our models are trained on publicly available and ethically sourced data, potential risks include unintended memorization or amplification of biases present in the training corpora. Moreover, while our work improves accessibility to speech foundation models (SFMs), downstream misuse–such as transcribing or translating copyrighted content–remains a concern and warrants further safeguards deployment in future efforts.

References

Md Mahfuz Ibn Alam and Antonios Anastasopoulos. 2024. A case study on filtering for end-to-end speech translation. *arXiv preprint arXiv:2402.01945*.

R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, et al. 2020. Common voice: A massively-multilingual speech corpus. In *Proc. LREC*.

Simone Balloccu, Patrícia Schmidtová, Mateusz Lango, and Ondrej Dusek. 2024. Leak, cheat, repeat: Data contamination and evaluation malpractices in closed-source LLMs. In *Proc. EACL*.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *arXiv preprint arXiv:2308.11596*.

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. 2023. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of ACL*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, et al. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proc. ICML*, volume 202.

BigScience Workshop et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Henry Chesbrough. 2015. From open science to open innovation. *Institute for Innovation and Knowledge Management, ESADE*.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, et al. 2023. Fleurs: Few-shot learning evaluation of universal representations of speech. In *Proc. SLT*.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuolin Yang, Zihan Liu, et al. 2024. Nvlm: Open frontier-class multimodal llms. *arXiv preprint arXiv:2409.11402*.

Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, et al. 2024. Molmo and pixmo: Open weights and open data for state-of-the-art multimodal models. *arXiv preprint arXiv:2409.17146*.

Yihong Dong, Xue Jiang, Huanyu Liu, Zhi Jin, Bin Gu, et al. 2024. Generalization or memorization: Data contamination and trustworthy evaluation for large language models. In *ACL Findings*.

Marco Gaido, Sara Papi, Luisa Bentivogli, Alessio Brutti, Mauro Cettolo, et al. 2024a. MOSEL: 950,000 hours of speech data for open-source speech foundation model training on EU languages. In *Proc. EMNLP*.

Marco Gaido, Sara Papi, Matteo Negri, and Luisa Bentivogli. 2024b. Speech translation with speech foundation models and large language models: What is there and what is missing? In *Proc. ACL*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proc. ICML*.

Dirk Groeneveld, Iz Beltagy, Evan Walsh, Akshita Bhagia, Rodney Kinney, et al. 2024. OLMo: Accelerating the science of language models. In *Proc. ACL*.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, et al. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech*.

Sudipta Kar, Giuseppe Castellucci, Simone Filice, Shervin Malmasi, and Oleg Rokhlenko. 2022. Preventing catastrophic forgetting in continual learning of new natural language tasks. In *Proc. ACM SIGKDD*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. EMNLP: System Demo.*

Sneha Kudugunta, Isaac Rayburn Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, et al. 2023. MADLAD-400: A multilingual and document-level large audited dataset. In *Proc. NeurIPS*.

Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, et al. 2023. Llm360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M Guerreiro, Ricardo Rei, Duarte M Alves, et al. 2024. Eurollm: Multilingual language models for europe. *arXiv preprint arXiv:2409.16235*.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *Proc. ICASSP*.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, et al. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech*.

Yifan Peng, Jinchuan Tian, William Chen, Siddhant Arora, Brian Yan, et al. 2024. Owsm v3.1: Better and faster open whisper-style speech models based on e-branchformer. In *Proc. Interspeech*.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, et al. 2023. Reproducing whisper-style training using an open-source toolkit and publicly available data. In *Proc. ASRU*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. Interspeech*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proc. ICML*.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proc. EMNLP*.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, et al. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proc. ACL*.

Qiang Sun, Yuanyi Luo, Sirui Li, Wenxiao Zhang, and Wei Liu. 2024. OpenOmni: A collaborative open source tool for building future-ready multimodal conversational agents. In *Proc. EMNLP: System Demo.*

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

Ioannis Tsiamas, Gerard I. Gállego, José A. R. Fonollosa, and Marta R. Costa-jussà. 2022. Shas: Approaching optimal segmentation for end-to-end speech translation. In *Proc. Interspeech*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, et al. 2017. Attention is all you need. In *Proc. NeurIPS*, volume 30.

Ruben Vicente-Saez and Clara Martinez-Fuentes. 2018. Open science now: A systematic literature review for an integrated definition. *Journal of Business Research*, 88.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, et al. 2021a. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proc. ACL*.

Changhan Wang, Yun Tang, Xutai Ma, Anne Wu, Dmytro Okhonko, and Juan Pino. 2020. fairseq s2t: Fast speech-to-text modeling with fairseq. In *Proc. AACL: System Demo.*

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021b. CoVoST 2 and Massively Multilingual Speech Translation. In *Proc. Interspeech*.

Matt White, Ibrahim Haddad, Cailean Osborne, Ahmed Abdelmonsef, Sachin Varghese, et al. 2024. The model openness framework: Promoting completeness and openness for reproducibility, transparency and usability in ai. *arXiv preprint arXiv:2403.13784*.

Brian Yan, Siddharth Dalmia, Yosuke Higuchi, Graham Neubig, Florian Metze, et al. 2023. CTC alignments improve autoregressive translation. In *Proc. EACL*.

A Training Data

The list of publicly available datasets is presented in Table 3.

Dataset	#hours		Label
	en	*it*	
CommonVoice v18 (Ardila et al., 2020)	1746	250	G
CoVoST2 (Wang et al., 2021b)	420	28	G
FLEURS (Conneau et al., 2023)	7	9	G
LibriSpeech (Panayotov et al., 2015)	358	-	G
MOSEL (Gaido et al., 2024a)	66,301	21,775	A
MLS (Pratap et al., 2020)	44,600	247	G
VoxPopuli-ASR (Wang et al., 2021a)	519	74	G
Total	113,951	22,383	G+A

Table 3: List of the publicly available training speech data for English (en) and Italian (it). "G" stands for gold labels while "A" for automatically generated labels.

To create the new psuedolabels, we leveraged the speech content of YouTube-Commons,[8] a dataset collecting YouTube videos released with the permissive 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 (Team, 2024), a lightweight VAD having low computational requirements. Lastly, the audio is split using SHAS (Tsiamas et al., 2022) to obtain segments suitable for training of around 16 seconds on average. The resulting dataset contains automatic transcripts, which we created with Whisper `large-v3`,[9] for 14,200k hours of speech for English (*en*) and 1,828k for Italian (*it*). Including data in Table 3, 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` (Kudugunta et al., 2023).[10] Following (Alam and Anastasopoulos, 2024), 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 ($r_{\text{min}} = 0.75, r_{\text{max}} = 1.45$ for en-it, and $r_{\text{min}} = 0.65, r_{\text{max}} = 1.35$ for it-en), resulting into 1.24% of data filtering for en-it and 3.08% for it-en. The final training set comprises the automatically translated speech data and the gold CoVoST2 and FLEURS datasets, resulting in 149,564 hours for *en-it* and 24,211 hours for *it-en*.

[8] https://hf.co/datasets/PleIAs/YouTube-Commons
[9] https://hf.co/openai/whisper-large-v3
[10] https://hf.co/google/madlad400-3b-mt

B FAMA Model Architecture

FAMA models come in two sizes: `small` and `medium`. FAMA-small has 12 Conformer encoder layers and 6 Transformer decoder layers. FAMA-medium has 24 encoder layers and 12 decoder layers. Each layer has 16 attention heads, an embedding dimension of 1,024, and a FFN dimension of 4,096. The Conformer encoder is preceded by two 1D convolutional layers with stride 2 and kernel size 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 (Kudo and Richardson, 2018) 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.

Detailed architectural settings are summarized in Table 4.

Encoder		
Component	*Size*	
	`small`	`medium`
Layer type	Conformer	
Number of layers	12	24
Attention heads	16	
Embedding dimension	1,024	
FFN dimension	4,096	
Convolutional Module kernel size	31	
Decoder		
Component	*Size*	
	`small`	`medium`
Layer type	Transformer	
Number of layers	6	12
Attention heads	16	
Embedding dimension	1,024	
FFN dimension	4,096	

Table 4: Architectural parameters of FAMA models.

C Training and Evaluation Procedures

FAMA Training Process. We train both models using a combination of three losses. First, a

label-smoothed cross-entropy loss ($\mathcal{L}_{\text{CE}}$) is applied to the decoder output, using the target text as the reference (transcripts for ASR and translations for ST). Second, a CTC loss (Graves et al., 2006) is computed using transcripts as reference ($\mathcal{L}_{\text{CTCsrc}}$) on the output of the 8$^{\text{th}}$ encoder layer for `small` and the 16$^{\text{th}}$ for `medium`. Third, a CTC loss on the final encoder output ($\mathcal{L}_{\text{CTCtgt}}$) is applied to predict the target text. The final loss is the weighted sum of the above-mentioned losses:

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

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 $\sim$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 (Vaswani et al., 2017) with a peak of 2e-3 and 25,000 warm-up steps. To cope with convergence issues similar to (Peng et al., 2024), 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 §3. 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 (Park et al., 2019) 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. All trainings are done using fairseq-S2T (Wang et al., 2020).

FAMA Evaluation. 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 (Yan et al., 2023) leveraging the CTC on the encoder output with weight 0.2. Inference is done using fairseq-S2T (Wang et al., 2020).

Terms of Comparison Evaluation. To ensure a fair comparison, we perform the inference with HuggingFace transformers[11] 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[12] inference code is used with beam size 3.[13]

Computational Time Benchmarking. The test set used for the computational time evaluation is CommonVoice on both *en* and *it* with a total duration of, respectively, 26.9 and 26.4 hours. The benchmarking is done on a single NVIDIA A40 40GB.

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

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

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