MOSEL: 950,000 Hours of Speech Data for Open-Source Speech Foundation Model Training on EU Languages

Marco Gaido, **Sara Papi**, **Luisa Bentivogli, Alessio Brutti, Mauro Cettolo, Roberto Gretter, Marco Matassoni, Mohamed Nabih, Matteo Negri**
Fondazione Bruno Kessler, Italy
{mgaido,spapi,bentivo,brutti,cettolo,gretter,matasso,mnabih,negri}@fbk.eu

Abstract

The rise of foundation models (FMs), coupled with regulatory efforts addressing their risks and impacts, has sparked significant interest in open-source models. However, existing speech FMs (SFMs) fall short of full compliance with the open-source principles, even if claimed otherwise, as no existing SFM has model weights, code, and training data publicly available under open-source terms. In this work, we take the first step toward filling this gap by focusing on the 24 official languages of the European Union (EU). We collect suitable training data by surveying automatic speech recognition datasets and unlabeled speech corpora under open-source compliant licenses, for a total of 950k hours. Additionally, we release automatic transcripts for 441k hours of unlabeled data under the permissive CC-BY license, thereby facilitating the creation of open-source SFMs for the EU languages.

github.com/hlt-mt/mosel
hf.co/datasets/FBK-MT/mosel

1 Introduction

The introduction of foundation models trained on large datasets is revolutionizing the landscape of many NLP fields (Bommasani et al., 2021), particularly with the release of Large Language Models (LLMs) that demonstrated impressive abilities on various tasks (Radford et al., 2019). The interest attracted by such models has come together with concerns about their risks and impact, as well as requests for a better understanding of their inner workings. On the one hand, this has led to regulatory efforts (European Parliament, 2023; Roberts et al., 2024), while on the other hand, it has sparked a growing interest in open-source models (Workshop et al., 2023; Groeneveld et al., 2024) that can be accessed and studied by anyone. However, it has been acknowledged that the term "*open source*" has been abused (Eiras et al., 2024; Liesenfeld and Dingemanse, 2024), being associated with any model whose weights are free to access (e.g., Touvron et al., 2023; Chiang et al., 2023), which is not sufficient to define a model as open source (OS).

In line with the Open Source Definition and its principles,[1] the Open Source Initiative defines as Open Source AI a "*system made available under terms that grant the freedoms to:* ***use*** *the system for any purpose without having to ask for permission*", "***study***", "***modify*** *[...] for any purpose*", and "***share*** *[...] with or without modifications, for any purpose*".[2] Specifically, it requires that the model and the code "*used to train and run the system*" are available under an OS license,[3] and that the training data is available under an OS-compliant license (White et al., 2024). This means that an OS model should not be trained on data released under licenses that restrict any of the four essential rights – use, study, modify, and share – for any purpose, including commercial use. Examples of OS-compliant licenses include CDLA-Permissive-2.0 and CC-BY-4.0, which only requires attribution (i.e., acknowledging the source or resource used). Instead, data released under licenses like CC-NC-4.0, which prohibits commercial use, or CC-SA-4.0, which mandates that derivative works have to be distributed under the same terms (thereby limiting the freedom to modify and share for any purpose), are not OS compliant.

Focusing on speech foundation models (SFMs), none of the existing ones complies with this definition. For instance, SeamlessM4T's model (Communication et al., 2023) is released under a license that is not OS compliant, while Whisper's model

Equal contribution.

[1] https://opensource.org/osd
[2] https://opensource.org/deepdive/drafts/the-open-source-ai-definition-draft-v-0-0-8
[3] https://opensource.org/licenses

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 13934–13947
November 12-16, 2024 ©2024 Association for Computational Linguistics

and inference code (Radford et al., 2023) are, but the training code and data are not public. Lastly, OWSM (Peng et al., 2023), although fulfilling most of the requirements, has been trained using datasets such as MuST-C (Di Gangi et al., 2019) and SPGISpeech (O'Neill et al., 2021), which have licenses that do not permit derivative works or commercial use. As a consequence, to the best of our knowledge, no current SFM satisfies the Open Source AI definition and can hence claim to be an Open Source SFM (OSSFM).

Considering the 24 official languages of the European Union (EU),[4] we take the first step towards filling this gap and, in particular, toward the creation of an EU-OSSFM: the collection of OS-compliant training data. With this aim, we survey the automatic speech recognition (ASR) datasets and the unlabeled speech corpora available for EU languages and list those that can be used to train an EU-OSSFM, for a total of 950k hours. This inventory of OS-compliant data, which will be continuously updated, is called **MOSEL** (Massive Open-source compliant Speech data for the European Languages) and is publicly available as a GitHub repository at: `github.com/hlt-mt/mosel`. In addition, to further ease the development of an EU-OSSFM, we automatically generated transcripts (i.e., pseudo-labels) for 441k hours of unlabeled data, which we release under the OS-compliant CC-BY 4.0 license on HuggingFace at: `hf.co/datasets/FBK-MT/mosel`. We conclude our work with an experiment on Maltese, one of the lowest-resourced languages, showing that the data can effectively be used for training ASR models.

2 Open Source Compliant Speech Data

This section surveys the available corpora that are admissible for developing an OSSFM for all 24 official EU languages. Accordingly, we include datasets that are freely accessible (i.e., excluding paid datasets) and whose data is released under an OS-compliant license (i.e., without restrictions on creating and redistributing derivative artifacts, including AI models).[5] This means that, in the case of the widespread Creative Commons (CC) licenses, we cannot include data released with non-derivative (ND), non-commercial (NC), or share-alike (SA)[6] restrictions. We also exclude datasets whose license is OS compliant but containing data released under a non-OS-compliant license. In fact, CC licenses "*allow licensed material to be included in collections* [...]*, however this does not change the license applicable to the original material*".[7]

In line with this indication, in cases where the transcripts are OS compliant (e.g., CC-BY where only attribution is required) but the corresponding speech (or part of it) is not, we document the dataset under the most restrictive license. For instance, GigaSpeech (Chen et al., 2021), which is released under Apache 2.0,[8] is categorized as non-OS compliant since it contains YouTube videos under restrictive CC licenses.[9] Similarly, MaSS (Zanon Boito et al., 2020) and CMU Wilderness (Black, 2019) are regarded as non-OS compliant since they are derived from the *Bible.is* data of the *Faith Comes By Hearing* organization with NC and ND terms of use.[10]

Table 1 lists the OS-compliant datasets with their license, number of hours, supported languages,[11] and whether they also contain transcripts.[12] The resulting **MOSEL** collection comprises 18 datasets, 7 of which are either in the Public Domain – i.e., without copyright terms (Fishman, 2006) – or licensed under CC-0, the most permissive CC license.[13] Overall, there are 505,7k hours of labeled data (i.e., including the transcripts). However, 87% of it comes from the YouTube-Commons dataset (PleIAs, 2024), for which manual inspection revealed some issues, as *i)* it includes videos without speech (e.g., with only music), *ii)* the language identification (LID) tag and the transcripts are often inaccurate, and *iii)* sentence-level segmentation of the speech is not provided (it contains unsegmented transcripts for

[4] `https://european-union.europa.eu/principles-countries-history/languages_en`

[5] `https://creativecommons.org/faq/#artificial-intelligence-and-cc-licenses`

[6] As the license of the resulting model "*must be a Creative Commons license with the same License Elements [...] or a BY-SA Compatible License*" (`https://creativecommons.org/licenses/by-sa/4.0/legalcode#s3b`), which is not compliant with open source terms.

[7] `https://creativecommons.org/faq/#if-i-create-a-collection-that-includes-a-work-offered-under-a-cc-license-which-licenses-may-i-choose-for-the-collection`

[8] `https://apache.org/licenses/LICENSE-2.0`

[9] `https://www.youtube.com/static?template=terms`

[10] `https://www.faithcomesbyhearing.com/terms`

[11] Represented as two-letter ISO 639 codes: `https://en.wikipedia.org/wiki/List_of_ISO_639_language_codes`.

[12] For completeness, in Appendix C, we list the most popular non-OS-compliant datasets, divided into those licensed under SA (Table 8), and under NC, ND, and other licenses (Table 9).

[13] `https://creativecommons.org/share-your-work/cclicenses/`

Name	License	Hours	Languages	Label
CommonVoice (Ardila et al., 2020)	CC-0	6,732	bg, cs, da, nl, en, et, fi, fr, de, el, hu, ga, it, lv, lt, mt, pl, pt, ro, sk, sl, es, sv	✅
CoVoST2 (Wang et al., 2021b)	CC-0	687	en, fr, it, es, pt, et, nl, sv, lv, sl	✅
CSS10 (Park and Mulc, 2019)	Public Domain	99	nl, fi, fr, de, el, hu, es	✅
EMU (Marasek et al., 2015)	CC-BY 3.0	56	pl	✅
EU Parliament (Chmiel et al., 2021)	CC-BY 4.0	32	pl	✅
FLEURS (Conneau et al., 2023)	CC-BY 4.0	215	bg, cs, da, nl, en, et, fi, fr, de, el, hu, ga, it, lv, lt, mt, pl, pt, ro, sk, sl, es, sv	✅
Large Corpus of Czech Parliament Plenary Hearings (Kratochvil et al., 2020)	CC-BY 4.0	444	cs	✅
LibriLight (Kahn et al., 2020)	Public Domain	57,706	en	❌
LibriTTS (Zen et al., 2019)	CC-BY 4.0	585	en	✅
LibriSpeech (Panayotov et al., 2015)	CC-BY 4.0	360	en	✅
LibriVoxDeEn (Beilharz et al., 2020)	Public Domain	547	de	✅
MC Speech (Czyżnikiewicz, 2022)	CC-0	22	pl	✅
MLS (Multilingual LibriSpeech) (Pratap et al., 2020)	CC-BY 4.0	50,687	nl, en, fr, de, it, pl, pt, es	✅
SIWIS (Honnet et al., 2017)	CC-BY 4.0	11	fr	✅
Speech Commands (Warden, 2018)	CC-BY 4.0	18	en	✅
VCTK (Yamagishi et al., 2019)	CC-BY 4.0	44	en	✅
VoxPopuli (Wang et al., 2021a)	CC-0	383,500	bg, hr, cs, da, nl, en, et, fi, fr, de, el, hu, it, lv, lt, mt, pl, pt, ro, sk, sl, es, sv	❌
		1,791	hr, cs, nl, en, et, fu, fr, de, hu, it, lt, pl, ro, sk, sl, es	✅
YouTube-Commons (PleIAs, 2024)	CC-BY 4.0	3,261	bg, cs, nl, en, et, fr, de, el, hu, it, pl, pt, ro, es	❌
		443,396	bg, cs, nl, en, et, fi, fr, de, el, hu, it, lv, lt, pl, pt, ro, es, sv	✅

Table 1: **MOSEL**🍇 speech datasets with OS-compliant license. We also report the total number of hours (*Hours*), languages supported (*Languages*), and whether they include reference transcripts (*Label*).

the entire YouTube videos). Therefore, further checks and processing work would be needed to effectively exploit the dataset for OSSFM training.

The total speech content (both labeled and unlabeled) amounts to 950,2k hours, which significantly exceeds the total data used to train most of the current SFMs (e.g., 680k hours for Whisper v2, 180k for OWSM), with the only exception of Whisper v3 whose training data comprises 5 million of hours. Even excluding the 446k hours of YouTube-Commons, the amount of data remains comparable, especially since Whisper v2 and OWSM target 99 and 151 languages respectively, instead of the 24 required for an EU-OSSFM.

Looking at language coverage, Table 2 shows that labeled data distribution is highly skewed towards English (see also Figure 1a in Appendix A.1). Indeed, only 6 other languages (de, es, fr it, nl, pt) can be considered as high-resource, with more than 3k hours. Instead, the unlabeled data is more evenly distributed (see also Figure 1b in Appendix A.1) and includes at least 8k hours for all EU languages but Irish, for which, unfortunately, we did not find unlabeled OS-compliant data.

3 Pseudo-labeling Process

The statistics reported in §2 highlight the importance of leveraging unlabeled data for training an OSSFM, given the scarcity of labeled material for most languages. When unlabeled data is available for model training, a common strategy consists of creating weak supervision (Zhou, 2017; Jia et al., 2019; Oramas et al., 2021; Zhang et al., 2022; Ren et al., 2023), which, in the context

Language	Label.	Unlabel.	Total
Bulgarian (bg)	111	17,609	17,720
Croatian (hr)	55	8,106	8,161
Czech (cs)	591	18,705	19,296
Danish (da)	20	13,600	13,620
Dutch (nl)	3,395	19,014	22,409
English (en)	437,239	84,704	521,943
Estonian (et)	60	10,604	10,664
Finnish (fi)	64	14,200	14,264
French (fr)	26,984	22,896	49,880
German (de)	9,236	23,228	32,464
Greek (el)	35	17,703	17,738
Hungarian (hu)	189	17,701	17,890
Irish (ga)	17	0	17
Italian (it)	3,756	21,933	25,689
Latvian (lv)	173	13,100	13,273
Lithuanian (lt)	36	14,400	14,436
Maltese (mt)	19	9,100	9,119
Polish (pl)	510	21,207	21,717
Portuguese (pt)	5,492	17,526	23,018
Romanian (ro)	121	17,906	18,021
Slovak (sk)	61	12,100	12,161
Slovenian (sl)	32	11,300	11,332
Spanish (es)	17,471	21,526	38,997
Swedish (sv)	58	16,300	16,358
Total	505,725	444,467	950,192

Table 2: **MOSEL** number of hours of labeled and unlabeled speech data for each official EU language.

of ASR, entails generating automatic transcripts. In light of the high computational resources demanded by this process for large-scale SFM training data, avoiding duplicated efforts across different institutions can significantly reduce the overall environmental impact and costs (Strubell et al., 2019), in line with Green AI principles (Schwartz et al., 2019). For this reason, we complement our inventory by providing practitioners with automatic transcripts for 441k hours of unlabeled speech coming from VoxPopuli and LibriLight.[14] The resulting pseudo-labeled data, whose statistics per language are presented in §A.2, covers nearly half of the total data available for training an EU-OSSFM and 23 of the 24 EU languages. In line with the spirit of this work, the transcripts are released under the OS-compliant CC-BY license at `hf.co/datasets/FBK-MT/mosel`.

The data is transcribed using Whisper large v3[15], which is released under the OS Apache 2.0 License that allows the generated content to be released under any license. In Appendix D, we report the ASR quality of Whisper across the EU languages. The inference is realized by feeding Whisper with the corresponding language ID and the <|notimestamp|> token, with 5 as beam size. As LibriLight, differently from VoxPopuli, contains segments longer than Whisper's maximum duration limit of 30s, we split them into chunks of up to 30s each. To ensure reproducibility, we will release the code under the Apache 2.0 Licence.

Costs. We executed all the inferences on NVIDIA A100 64GB GPUs, on which we managed to fit 16 samples per batch and enabled FlashAttention (Dao et al., 2022) to speed up the generation process. In this way, we reached a throughput of ~1.5-2k samples per GPU hour. As a result, the transcription process required a total of ~25,500 GPU hours. On popular cloud services such as AWS, this would translate to >100k USD[16] and 35,625 kgCO_2eq estimated emissions.[17]

[14] YouTube-Commons was excluded due to the issues described in §2.

[15] `https://huggingface.co/openai/whisper-large-v3` with HuggingFace v4.38.2.

Model	CommonVoice	FLEURS
Whisper large v3	80.8	73.8
label. + pseudo-lab.	39.4	38.9
label. + *filtered* pseudo-lab.	23.8	24.5

Table 3: ASR results (WER↓) for Maltese. We compare Whisper and our models trained respectively *i)* on labeled and pseudo-labeled **MOSEL** data and *ii)* on the same data with filters applied to pseudo-labeled data.

4 Proof of Concept on Maltese

To showcase that the datasets collected in our survey (§2) and the generated transcripts (§3) constitute suitable training data for an EU-OSSFM, we conduct a proof-of-concept experiment on Maltese. Maltese was chosen because it is *i)* one of the lowest-resourced languages, and *ii)* the one for which Whisper achieves the worst results, as shown in Appendix D.[18]

For our experiments, we first attempted to train an ASR model using only supervised data, but it failed to converge due to its limited size (16 hours). Therefore, we trained a model using the few labeled data together with the pseudo-labeled data.[19] As an additional investigation, we also applied to the pseudo-labeled data simple filtering methods to remove audios containing other languages and automatic transcripts containing hallucinations (see Appendix B.2). Results presented in Table 3 show

[16] As of June 10th 2024, 8 A100 GPUs cost >32 USD. See `https://aws.amazon.com/it/ec2/instance-types/p4/`.

[17] Estimations were conducted using the MachineLearning Impact calculator presented in (Lacoste et al., 2019).

[18] With the only exception of Irish, which has only 17 collected hours and is not even supported by Whisper.

[19] For full experimental details see Appendix B.

that the model trained with all data doubles the performance of Whisper (∼39 vs. ∼80 WER). Considering the very low performance of Whisper, which was used to create the automatic transcripts, the contribution of the pseudo-labeled data is noticeable. Also interesting is the further improvement obtained when unlabeled data are filtered (∼24 WER). These experiments support the conclusion that the collected and transcribed data represent a promising bedrock for developing an EU-OSSFM.

5 Conclusions

In response to the urgent need for truly open-source foundation models, this work takes the first step toward an EU open-source speech foundation model, which is the collection of suitable training data called **MOSEL**. To this end, we first surveyed the labeled and unlabeled speech datasets for automatic speech recognition that feature at least one of the 24 official EU languages and are available under a license compliant with the open-source terms. We then complemented this effort with the creation and release of automatic transcripts for the available unlabeled data. Overall, we collected more than 950k hours of speech content suitable for the training of an EU open-source speech foundation model, also demonstrating its usefulness in Maltese, one of the lowest-resourced languages.

6 Acknowledgments

The work presented in this paper is funded by the European Union's Horizon research and innovation programme under grant agreement No 101135798, project Meetween (My Personal AI Mediator for Virtual MEETtings BetWEEN People) and the PNRR project FAIR - Future AI Research (PE00000013), under the NRRP MUR program funded by the NextGenerationEU. We also acknowledge the CINECA award (MAGIS) under the ISCRA initiative, for the availability of high-performance computing resources and support.

7 Limitations

Collecting Open Irish Data. An important future direction to expand this work is represented by collecting and releasing new material – possibly with human-generated transcripts – under permissive licenses for the least-resourced language. This is especially critical for Irish, for which we were able to collect only 17 hours of (labeled) speech.

Data Curation of Available Resources. As noted in §2, the quality of the supervision of the surveyed dataset cannot always be taken for granted, advocating for dedicated inspections before using it to train an OSSFM. This is particularly true for the metadata and transcripts of YouTube videos under OS-compliant licenses as those collected in YouTube-Commons.

Quality of Pseudo-labels and Filtering Techniques. The quality of Whisper outputs greatly varies across the 24 languages. In particular, the WER of Whisper for Maltese is high (80.8 on the CommonVoice test set and 73.8 on the FLEURS test set). As such, filtering strategies aiming at identifying unreliable transcriptions may be required for the successful training of OSSFM, especially for low-resource languages. Indeed, as already seen in §4, even simple filtering techniques proved to be effective in greatly improving ASR performance. More advanced filtering techniques can provide further benefits for the quality of the resulting model. However, data cleaning and normalization are common steps in training pipelines, going beyond the scope of this work.

Beyond EU languages. This paper has focused only on the 24 EU languages. An obvious next step for this work is its extension to many other spoken languages, with the final goal of covering hundreds of languages and leading to the creation of a universal OSSFM.

References

Mohamed Anwar, Bowen Shi, Vedanuj Goswami, Wei-Ning Hsu, Juan Pino, and Changhan Wang. 2023. Muavic: A multilingual audio-visual corpus for robust speech recognition and robust speech-to-text translation. *arXiv preprint arXiv:2303.00628*.

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

Benjamin Beilharz, Xin Sun, Sariya Karimova, and Stefan Riezler. 2020. LibriVoxDeEn: A corpus for German-to-English speech translation and German speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3590–3594, Marseille, France. European Language Resources Association.

Alan W Black. 2019. CMU wilderness multilingual speech dataset. In *ICASSP 2019 - 2019 IEEE Inter-*

national Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5971–5975.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, et al. 2021. On the Opportunities and Risks of Foundation Models. *ArXiv*.

Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre Wellner. 2006. The AMI Meeting Corpus: A Pre-announcement. In *Machine Learning for Multimodal Interaction*, pages 28–39, Berlin, Heidelberg. Springer Berlin Heidelberg.

Guoguo Chen, Shuzhou Chai, Guan-Bo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, Mingjie Jin, Sanjeev Khudanpur, Shinji Watanabe, Shuaijiang Zhao, Wei Zou, Xiangang Li, Xuchen Yao, Yongqing Wang, Zhao You, and Zhiyong Yan. 2021. GigaSpeech: An Evolving, Multi-Domain ASR Corpus with 10,000 Hours of Transcribed Audio. In *Proc. Interspeech 2021*, pages 3670–3674.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality.

Agnieszka Chmiel, Przemysław Janikowski, Marta Kajzer-Wietrzny, Danijel Koržinek, and Dariusz Jakubowski. 2021. EU parliament speech corpus. CLARIN-PL digital repository.

Seamless Communication et al. 2023. SeamlessM4T: Massively Multilingual & Multimodal Machine Translation. *Preprint*, arXiv:2308.11596.

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

Mateusz Czyżnikiewicz. 2022. *Analiza porównawcza korpusów nagrań mowy dla celów syntezy mowy w języku polskim*. Master's thesis, Warsaw University of Technology. Available at `http://dx.doi.org/10.13140/RG.2.2.26293.24800`.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 35:16344–16359.

Isin Demirsahin, Oddur Kjartansson, Alexander Gutkin, and Clara Rivera. 2020. Open-source Multi-speaker Corpora of the English Accents in the British Isles. In *Proceedings of The 12th Language Resources and Evaluation Conference (LREC)*, pages 6532–6541, Marseille, France. European Language Resources Association (ELRA).

Mattia A. Di Gangi, Roldano Cattoni, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2019. MuST-C: a Multilingual Speech Translation Corpus. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2012–2017, Minneapolis, Minnesota. Association for Computational Linguistics.

Francisco Eiras, Aleksandar Petrov, Bertie Vidgen, Christian Schroeder, Fabio Pizzati, Katherine Elkins, Supratik Mukhopadhyay, Adel Bibi, Aaron Purewal, Csaba Botos, Fabro Steibel, Fazel Keshtkar, Fazl Barez, Genevieve Smith, Gianluca Guadagni, Jon Chun, Jordi Cabot, Joseph Imperial, Juan Arturo Nolazco, Lori Landay, Matthew Jackson, Phillip H. S. Torr, Trevor Darrell, Yong Lee, and Jakob Foerster. 2024. Risks and Opportunities of Open-Source Generative AI. *Preprint*, arXiv:2405.08597.

European Parliament. 2023. EU AI Act: first regulation on artificial intelligence.

Soline Felice, Solène Evain, Solange Rossato, and François Portet. 2024. Audiocite.net: A large spoken read dataset in french. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*.

S. Fishman. 2006. *The Public Domain: How to Find & Use Copyright-free Writings, Music, Art & More*. Public Domain: How to Find and Use Copyright-Free Writings, Music, Art& More Series. Nolo.

Daniel Galvez, Greg Diamos, Juan Manuel Ciro Torres, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Max Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, page 369–376, New York, NY, USA.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Raghavi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E.

Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Emma Strubell, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Luke Zettlemoyer, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. OLMo: Accelerating the Science of Language Models. *Preprint*, arXiv:2402.00838.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, and Ruoming Pang. 2020. Conformer: Convolution-augmented Transformer for Speech Recognition. In *Proc. Interspeech 2020*, pages 5036–5040.

Jan Hajič, Petr Pajas, Pavel Ircing, Jan Romportl, Nino Peterek, Miroslav Spousta, Marie Mikulová, Martin Grůber, and Milan Legát. 2017. Prague DaTabase of spoken czech 1.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia Tomashenko, and Yannick Esteve. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer: 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18–22, 2018, Proceedings 20*, pages 198–208. Springer.

Carlos D. Hernandez-Mena. 2019. TEDx Spanish Corpus. Audio and transcripts in Spanish taken from the TEDx Talks; shared under the CC BY-NC-ND 4.0 license. Web Download.

Carlos Daniel Hernandez Mena, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke van der Plas, Amanda Muscat, and Ian Padovani. 2020. MASRI-HEADSET: A Maltese corpus for speech recognition. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6381–6388, Marseille, France. European Language Resources Association.

Pierre-Edouard Honnet, Alexandros Lazaridis, Philip N. Garner, and Junichi Yamagishi. 2017. The siwis french speech synthesis database ? design and recording of a high quality french database for speech synthesis.

J. Iranzo-Sánchez, J. A. Silvestre-Cerdà, J. Jorge, N. Roselló, A. Giménez, A. Sanchis, J. Civera, and A. Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8229–8233.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12).

Ye Jia, Melvin Johnson, Wolfgang Macherey, Ron J. Weiss, Yuan Cao, Chung-Cheng Chiu, Naveen Ari, Stella Laurenzo, and Yonghui Wu. 2019. Leveraging weakly supervised data to improve end-to-end speech-to-text translation. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7180–7184.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-Light: A Benchmark for ASR with Limited or No Supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. `https://github.com/facebookresearch/libri-light`.

Andreas Kirkedal, Marija Stepanović, and Barbara Plank. 2020. FT Speech: Danish Parliament Speech Corpus. In *Proc. Interspeech 2020*.

Arne Köhn, Florian Stegen, and Timo Baumann. 2016. Mining the spoken wikipedia for speech data and beyond. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).

Rostislav Kolobov, Olga Okhapkina, Andrey Platunov Olga Omelchishina, Roman Bedyakin, Vyacheslav Moshkin, Dmitry Menshikov, and Nikolay Mikhaylovskiy. 2021. Mediaspeech: Multilanguage asr benchmark and dataset. *Preprint*, arXiv:2103.16193.

Matěj Korvas, Ondřej Plátek, Ondřej Dušek, Lukáš Žilka, and Filip Jurčíček. 2014. Free English and Czech telephone speech corpus shared under the CC-BY-SA 3.0 license. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4423–4428, Reykjavik, Iceland. European Language Resources Association (ELRA).

Jonáš Kratochvil, Peter Polák, and Ondřej Bojar. 2020. Large corpus of Czech parliament plenary hearings. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6363–6367, Marseille, France. European Language Resources Association.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia.

Alexandre Lacoste, Alexandra Luccioni, Victor Schmidt, and Thomas Dandres. 2019. Quantifying the carbon emissions of machine learning. *arXiv preprint arXiv:1910.09700*.

Andreas Liesenfeld and Mark Dingemanse. 2024. Rethinking open source generative AI: open-washing and the EU AI Act. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*.

Nikola Ljubešić, Danijel Koržinek, Peter Rupnik, and Ivo-Pavao Jazbec. 2022. ParlaSpeech-HR - a freely available ASR dataset for Croatian bootstrapped from the ParlaMint corpus. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 111–116, Marseille, France. European Language Resources Association.

Krzysztof Marasek, Danijel Koržinek, Łukasz Brocki, and Kamila Jankowska-Lorek. 2015. Clarin-PL studio corpus (EMU). CLARIN-PL digital repository.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Peter Mihajlik, Andras Balog, Tekla Etelka Graczi, Anna Kohari, Balázs Tarján, and Katalin Mady. 2022. BEA-base: A benchmark for ASR of spontaneous Hungarian. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1970–1977, Marseille, France. European Language Resources Association.

Sergio Oramas, Massimo Quadrana, and Fabien Gouyon. 2021. Bootstrapping a music voice assistant with weak supervision. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Industry Papers*, pages 49–55, Online. Association for Computational Linguistics.

Patrick K. O'Neill, Vitaly Lavrukhin, Somshubra Majumdar, Vahid Noroozi, Yuekai Zhang, Oleksii Kuchaiev, Jagadeesh Balam, Yuliya Dovzhenko, Keenan Freyberg, Michael D. Shulman, Boris Ginsburg, Shinji Watanabe, and Georg Kucsko. 2021. SPGISpeech: 5,000 Hours of Transcribed Financial Audio for Fully Formatted End-to-End Speech Recognition. In *Proc. Interspeech 2021*, pages 1434–1438.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: An ASR corpus based on public domain audio books. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210.

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

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

Kyubyong Park and Thomas Mulc. 2019. CSS10: A Collection of Single Speaker Speech Datasets for 10 Languages. In *Proc. Interspeech 2019*, pages 1566–1570.

Yifan Peng, Jinchuan Tian, Brian Yan, Dan Berrebbi, Xuankai Chang, Xinjian Li, Jiatong Shi, Siddhant Arora, William Chen, Roshan Sharma, Wangyou Zhang, Yui Sudo, Muhammad Shakeel, Jee-Weon Jung, Soumi Maiti, and Shinji Watanabe. 2023. Reproducing whisper-style training using an open-source toolkit and publicly available data. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Piotr Pęzik. 2018. Increasing the accessibility of time-aligned speech corpora with spokes mix. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Piotr Pęzik, Sylwia Karasińska, Anna Cichosz, Łukasz Jałowiecki, Konrad Kaczyński, Małgorzata Krawentek, Karolina Walkusz, Paweł Wilk, Mariusz Kleć, Krzysztof Szklanny, et al. 2023. SpokesBiz–an Open Corpus of Conversational Polish. *arXiv preprint arXiv:2312.12364*.

PleIAs. 2024. PleIAs/YouTube-Commons · Datasets at Hugging Face — huggingface.co. `https://huggingface.co/datasets/PleIAs/YouTube-Commons`. [Accessed 10-06-2024].

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. MLS: A Large-Scale Multilingual Dataset for Speech Research. In *Proc. Interspeech 2020*, pages 2757–2761.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning*, pages 28492–28518. PMLR.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Vipula Rawte, Swagata Chakraborty, Agnibh Pathak, Anubhav Sarkar, S.M Towhidul Islam Tonmoy, Aman Chadha, Amit Sheth, and Amitava Das. 2023. The troubling emergence of hallucination in large language models - an extensive definition, quantification, and prescriptive remediations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2541–2573, Singapore. Association for Computational Linguistics.

Zeyu Ren, Shuihua Wang, and Yudong Zhang. 2023. Weakly supervised machine learning. *CAAI Transactions on Intelligence Technology*, 8(3):549–580.

Huw Roberts, Emmie Hine, Mariarosaria Taddeo, and Luciano Floridi. 2024. Global AI governance: barriers and pathways forward. *International Affairs*, 100(3):1275–1286.

Elizabeth Salesky, Matthew Wiesner, Jacob Bremerman, Roldano Cattoni, Matteo Negri, Marco Turchi, Douglas W. Oard, and Matt Post. 2021. Multilingual TEDx Corpus for Speech Recognition and Translation. In *Proceedings of Interspeech*.

Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: a large-scale dataset for multimodal language understanding. In *Proceedings of the Workshop on Visually Grounded Interaction and Language (ViGIL)*. NeurIPS.

Roy Schwartz, Jesse Dodge, Noah A. Smith, and Oren Etzioni. 2019. Green ai. *Preprint*, arXiv:1907.10597.

Imdat Solak. 2019. The M-AILABS Speech Dataset. `https://www.caito.de/2019/01/03/the-m-ailabs-speech-dataset/`. [Accessed 04-06-2024].

Adriana Stan, Florina Dinescu, Cristina Tiple, Serban Meza, Bogdan Orza, Magdalena Chirila, and Mircea Giurgiu. 2017. The SWARA Speech Corpus: A Large Parallel Romanian Read Speech Dataset. In *Proceedings of the 9th Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, Bucharest, Romania.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Tatoeba. 2017. Tatoeba eng. `https://downloads.tatoeba.org/audio/`. [Accessed 04-06-2024].

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Jörgen Valk and Tanel Alumäe. 2021. VoxLingua107: a dataset for spoken language recognition. In *Proc. IEEE SLT Workshop*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30.

Darinka Verdonik, Andreja Bizjak, Mirjam Sepesy Maučec, Lucija Gril, Simon Dobrišek, Janez Križaj, Gregor Strle, Marko Bajec, Iztok Lebar Bajec, Tjaša Jelovšek, Jure Lokovšek, Mitja Trojar, Tomaž Erjavec, Mitja Bernjak, Jerneja Žganec Gros, Peter Čakš, Matevž Pucer, Mitja Cvetko, Jani Pavlič, Marijana Zelenik, Marija Ivanovska, Klemen Grm, Jure Longyka, Aleš Mihelič, Boštjan Vesnicer, and Naum Dretnik. 2023. ASR database ARTUR 1.0 (transcriptions). Slovenian language resource repository CLARIN.SI.

Darinka Verdonik, Iztok Kosem, Ana Zwitter Vitez, Simon Krek, and Marko Stabej. 2013. Compilation, transcription and usage of a reference speech corpus: The case of the slovene corpus gos. *Language Resources and Evaluation*, 47.

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

Changhan Wang, Anne Wu, Jiatao Gu, and Juan Pino. 2021b. CoVoST 2 and Massively Multilingual Speech Translation. In *Proc. Interspeech 2021*, pages 2247–2251.

Pete Warden. 2018. Speech commands: A dataset for limited-vocabulary speech recognition. *arXiv preprint arXiv:1804.03209*.

Matt White, Ibrahim Haddad, Cailean Osborne, XiaoYang Liu Yanglet, Ahmed Abdelmonsef, and Sachin Varghese. 2024. The model openness framework: Promoting completeness and openness for reproducibility, transparency, and usability in artificial intelligence. *Preprint*, arXiv:2403.13784.

BigScience Workshop et al. 2023. Bloom: A 176b-parameter open-access multilingual language model. *Preprint*, arXiv:2211.05100.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English Multi-speaker Corpus for CSTR Voice Cloning Toolkit (version 0.92).

Marcely Zanon Boito, William Havard, Mahault Garnerin, Éric Le Ferrand, and Laurent Besacier. 2020. MaSS: A large and clean multilingual corpus of sentence-aligned spoken utterances extracted from the Bible. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6486–6493, Marseille, France. European Language Resources Association.

Heiga Zen, Viet Dang, Rob Clark, Yu Zhang, Ron J. Weiss, Ye Jia, Zhifeng Chen, and Yonghui Wu. 2019. LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech. In *Proc. Interspeech 2019*, pages 1526–1530.

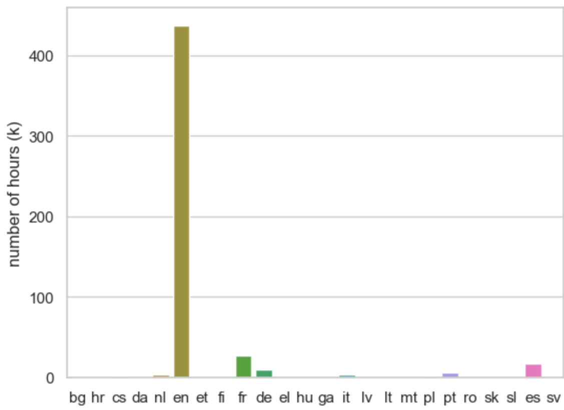
Yu Zhang, Daniel S. Park, Wei Han, James Qin, Anmol Gulati, Joel Shor, Aren Jansen, Yuanzhong Xu, Yanping Huang, Shibo Wang, Zongwei Zhou, Bo Li, Min Ma, William Chan, Jiahui Yu, Yongqiang Wang, Liangliang Cao, Khe Chai Sim, Bhuvana Ramabhadran, Tara N. Sainath, Françoise Beaufays, Zhifeng Chen, Quoc V. Le, Chung-Cheng Chiu, Ruoming Pang, and Yonghui Wu. 2022. BigSSL: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1519–1532.

Zhi-Hua Zhou. 2017. A brief introduction to weakly supervised learning. *National Science Review*, 5(1):44–53.

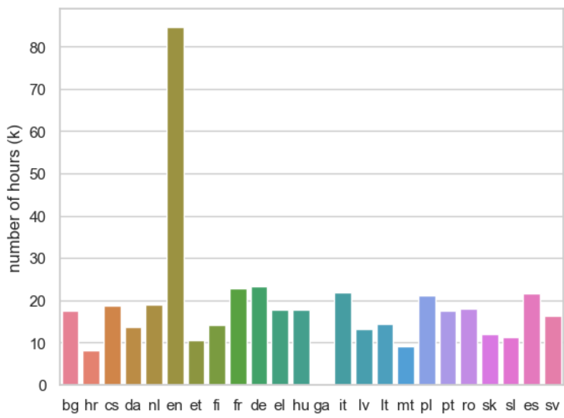
A Data Statistics

A.1 Labeled and Unlabeled Data Distribution

Data distributions for both labeled and unlabeled data discussed in §2 and referred to Table 1 are presented in, respectively, Figure 1a and 1b.



(a) Labeled



(b) Unlabeled

Figure 1: Labeled and unlabeled data distribution of the OS-compliant collected speech for each EU language.

A.2 Pseudo-labeled Data Statistics

The total number of hours of pseudo-labeled data described in §3 are shown in Table 4. The data distribution is similar to those of unlabeled data presented in §A.2 due to the nearly complete overlap with the retrieved unlabeled data, as already discussed in §3.

B Experimental Settings

B.1 Model and Training Settings

We train a sequence-to-sequence model whose encoder is a 12-layer Conformer (Gulati et al., 2020) and whose decoder is a 6-layer Transformer (Vaswani et al., 2017). The Conformer encoder is preceded by two 1D convolutional layers with

Language	Pseudo-labeled (hours)
Bulgarian (bg)	17,600
Croatian (hr)	8,100
Czech (cs)	18,700
Danish (da)	13,600
Dutch (nl)	19,000
English (en)	81,806
Estonian (et)	10,600
Finnish (fi)	14,200
French (fr)	22,800
German (de)	23,200
Greek (el)	17,700
Hungarian (hu)	17,700
Italian (it)	21,900
Latvian (lv)	13,100
Lithuanian (lt)	14,400
Maltese (mt)	9,100
Polish (pl)	21,200
Portuguese (pt)	17,500
Romanian (ro)	17,900
Slovak (sk)	12,100
Slovenian (sl)	11,300
Spanish (es)	21,400
Swedish (sv)	16,300
Total	441,206

Table 4: Number of hours for the pseudo-labeled data that we make available for each official EU language.

stride 2 and kernel size 5. We use an embedding size of 512 and an internal feed-forward dimension of 2048. The convolutional modules of the Conformer layers have a 31-feature kernel. The target vocabulary is built with size 8,000 using SentencePiece (Kudo, 2018), while the input audio is represented with 80 Mel-filterbank features extracted every 10 ms with a window of 25 ms. As a result, the model has 116M parameters in total.

We use label-smoothed cross-entropy loss on the decoder output (with 0.1 as label-smoothing factor), complemented with a CTC (Graves et al., 2006) loss (summed with 0.5 weight) trained on the output of the 8th encoder layers to facilitate the convergence of the model. The model was optimized with Adam ($\beta_1, \beta_2 = 0.9, 0.98$) using Noam learning rate scheduler (Vaswani et al., 2017) with 2e-3 as peak learning rate and 25,000 warmup steps. To avoid overfitting, we set dropout to 0.1 and weight decay to 0.001 and apply SpecAugment (Park et al., 2019) during training. To further ease the convergence of the model, we initialize the Conformer encoder weights with those of a similar ASR model trained on 4k hours of labeled English data, comprising CommonVoice, Librispeech, CoVoST, and VoxPopuli. We train the models with mini-batches of 40,0000 tokens and 2 as update frequency on 4 NVIDIA Ampere A100 GPUs (64GB RAM) for 150k updates and average the last 7 checkpoints.

Our experiments are conducted with the open-source repository available at https://github.com/hlt-mt/FBK-fairseq/ using the padding-safe implementation of the Conformer encoder (Papi et al., 2024). Results in Word Error Rate (WER) are computed using the Whisper Normalizer[20] and, then, JiWER[21] for computing the metric.

B.2 Data Filtering

B.2.1 LID

To check for possible inconsistencies between the metadata released in VoxPopuli and the actual content of speech segments, we check the actual spoken language with an automatic language identifier (LID). In fact, as in the transcription process described in §3 we force the language to the one provided in the metadata, these segments may be paired with noisy transcripts. The LID was carried out using the Whisper *large v3* model, as done for the transcription process, and it was performed by letting the model predict the language tag and taking the language with the highest probability.

LID	Portion (%)
mt	77.1
en	9.9
it	3.5
fr	2.2
ar	1.9
other	5.4

Table 5: Identified languages on the Maltese section of VoxPopuli (reported as %).

Table 5 shows the results. Upon a manual inspection, we noticed that the samples predicted as Maltese are indeed all correct. Similarly, the LID appeared mostly correct when predicting languages different from Maltese, except for the samples identified as Italian or Arabic which are sometimes Maltese speech. However, given the not-so-high amount of mislabeled data and to be on the safe side, in our experiments with "simple filters" we opted for filtering all the data recognized with a language different from Maltese, removing ∼23% of the 9k VoxPopuli hours.

To ensure the reproducibility of our experiments and to let practitioners leverage this information for their filtering strategies while creating OSSFM, we will release the LID output for all the transcribed unlabeled data under the CC-BY license.

[20] https://pypi.org/project/whisper-normalizer/
[21] https://pypi.org/project/jiwer/

id	Reference	Automatic Transcript
1	Where is the victim?	Yes, where's the victim?
2	Here	Hey, hey, hey, here, hey. No, no, no, no, no, no, no, no.
3	Good, and now to get hold of little Henny.	Shop for a moment, give Hennie a hand.

Table 6: Examples of hallucinations in Whisper outputs.

B.2.2 Textual Hallucinations

In the context of LLMs, hallucinations refer to "*the generation of content that deviates from the real facts, resulting in unfaithful outputs*" (Maynez et al., 2020; Rawte et al., 2023). In our context of ASR, they have been analogously defined as "*nonsensical, or unfaithful to the provided source input*" (Ji et al., 2023). Specifically, here we focus on the detection of *nonsensical* hallucinations, in which "*the generated text fails to convey any relevant or comprehensible information*",[22] while those related to semantic aspects are ignored.

Table 6 shows examples of hallucinated texts generated by Whisper in English. It can be noted that, in line 2, the word *here* is surrounded by many spurious "*hey,*" and that the successive sentence consists of a sequence of equally spurious "*no,*". This typically happens when background noise or music is present in the audio content, making the transcription task more difficult.

Another issue that can affect, although less frequently, the text generated by LLMs in general and by Whisper in particular, is the presence of very long and noisy strings like "T-J-N-D-F-Z-3-2-8-W-M-L-G-0-Z-P-[. . .]" and "Amen.Amen.Amen.Amen.Amen.Amen.[. . .]". Moreover, we noted that, for some languages, the decoding of entire audio segments sometimes generates one single, very common word, like "*Děkuji*" for Czech and "*Ačiū*" for Lithuanian, both corresponding to "*Thank you*". Although being correct in some cases, since for the most reliable languages (e.g., English and German) transcripts with a single word are relatively rare, we chose to consider this phenomenon as hallucination.

In conclusion, we decided to flag the segments containing all the above-described hallucinations, with the option of filtering them out during training. Also in this case, for the sake of reproducibility and to enable the adoption of similar approaches, we released the hallucination-detection metadata.

[22] https://masterofcode.com/blog/hallucinations-in-llms-what-you-need-to-know-before-integration

Language	CommonVoice	FLEURS
bg	14.3	12.5
hr	-	10.8
cs	9.0	10.1
da	18.1	12.0
nl	4.3	5.2
en	9.3	4.1
et	29.9	18.1
fi	24.6	7.7
fr	10.8	5.3
de	5.7	4.9
el	13.7	10.9
hu	13.4	12.9
ga	-	-
it	5.5	3.0
lv	16.7	19.4
lt	27.6	23.7
mt	80.8	73.8
pl	6.0	4.6
pt	5.9	4.1
ro	10.8	8.2
sk	23.4	9.2
sl	16.8	18.3
es	4.7	2.8
sv	8.3	7.6

Table 7: WER (↓) reported for Whisper large v3 (Radford et al., 2019) across the 24 European languages on CommonVoice and FLEURS.

C Non-open Datasets

C.1 CC-BY-SA

The collection of datasets with the SA license, which is not compliant with open-source criteria, is presented in Table 8.

C.2 CC-NC, -ND, and others

The collection of the most well-known datasets with a license that is not compliant with open-source criteria is presented in Table 9.

D Whisper Performance on EU Languages

Table 7 reports the WER scores obtained using Whisper on the 24 European languages. Maltese stands out as the worst language by a wide margin, with a very high WER (73.8 on FLEURS) indicating a limited ability to address the Maltese ASR task. All other languages display much lower WER, as only Estonian, Latvian, Lithuanian, and Slovenian exceed 15 WER, while high-resource

Name	License	hours	Languages	Label
ARTHUR 1.0 (Verdonik et al., 2023)	CC-BY-SA 4.0	884	sl	✅
Vystadial (Korvas et al., 2014)	CC-BY-SA 3.0	63	en, cs	✅
ParlaSpeech-HR (Ljubešić et al., 2022)	CC-BY-SA	1,816	hr	✅
People's Speech (Galvez et al., 2021)	CC-BY-SA 4.0	30,000	en	✅
SWC (Köhn et al., 2016)	CC-BY-SA 4.0	996	de, en, nl	❌
SWC-ASR (Köhn et al., 2016)	CC-BY-SA 4.0	510	de, en, nl	✅
UK and Ireland English Dialect (Demirsahin et al., 2020)	CC-BY-SA 4.0	31	ga	✅

Table 8: Speech datasets with Share-Alike (SA) license. If more languages are included, the sum is presented.

Name	License	hours (k)	Languages	Label
AMI (Carletta et al., 2006)	CC-BY-NC 4.0	100	en	✅
AudioCite.net (Felice et al., 2024)	CC-BY-NC	6,682	fr	✅
BEA-Base (Mihajlik et al., 2022)	NC	71	hu	✅
CMU Wilderness (Black, 2019)	NC, ND	236	en, fi, fr, pl, pt, to, es, sv	✅
Europarl-ST (Iranzo-Sánchez et al., 2020)	CC-BY-NC 4.0	201	en, fr, de, it, es, pt, pl, ro, nl	✅
FT Speech (Kirkedal et al., 2020)	NC	1,800	da	✅
GigaSpeech (Chen et al., 2021)	YouTube License	33,000	en	❌
GigaSpeech-ASR (Chen et al., 2021)	YouTube License	10,000	en	✅
GOS (Verdonik et al., 2013)	CC-BY-SA-NC 2.5	120	sl	✅
How-2 (Sanabria et al., 2018)	YouTube License	2,000	en	❌
How-2 ASR (Sanabria et al., 2018)	YouTube License	300	en	✅
M-AILABS (Solak, 2019)	Project Gutenberg License	867	en, fr, de, it, pl, es	✅
MASRI (Hernandez Mena et al., 2020)	NC	8	mt	✅
MaSS (Zanon Boito et al., 2020)	NC, ND	126	en, fi, fr, hu, ro, es	✅
MediaSpeech (Kolobov et al., 2021)	YouTube License	20	fr	✅
mTEDx (Salesky et al., 2021)	CC-BY-NC-ND 4.0	679	fr, el, it, pt, es	✅
MuAViC (Anwar et al., 2023)	CC-BY-NC 4.0	1,079	en, el, es, fr, it, pt	✅
MuST-C (Di Gangi et al., 2019)	CC-BY-NC-ND 4.0	504	en	✅
PDTSC1.0 (Hajič et al., 2017)	CC-BY-NC-SA 4.0	122	cs	✅
PELCRA (Pęzik, 2018)	CC-BY-NC	100	pl	✅
SpokesBiz (Pęzik et al., 2023)	CC-BY-NC-ND	650	pl	✅
SPGISpeech (O'Neill et al., 2021)	NC	5,000	en	✅
SWARA (Stan et al., 2017)	CC-BY-NC 4.0	21	ro	✅
Tatoeba ENG (Tatoeba, 2017)	CC-BY-NC-ND	200	en	✅
TEDLIUM v3 (Hernandez et al., 2018)	CC-BY-NC-ND 3.0	452	en	✅
TEDx Spanish (Hernandez-Mena, 2019)	CC-BY-NC-ND 4.0	24	es	✅
VoxLingua107 (Valk and Alumäe, 2021)	YouTube License	1,352	bg, hr, cs, da, nl, en, et, fi, fr, de, el, hu, it, lv, lt, mt, pl, pt, ro, sk, sl, es, sv	❌

Table 9: Non-open speech datasets. If more languages are included, the sum is presented.

languages such as Dutch, English, Italian, German, and Spanish consistently achieve WER close or lower than 5.