EuroSpeech: A Multilingual Speech Corpus

Samuel Pfisterer **Florian Grötschla** **Luca A. Lanzendörfer**
Florian Yan **Roger Wattenhofer**

ETH Zurich
{spfisterer, fgroetschla, lanzendoerfer, floyan, wattenhofer}@ethz.ch

Abstract

Recent progress in speech processing has highlighted that high-quality performance across languages requires substantial training data for each individual language. While existing multilingual datasets cover many languages, they often contain insufficient data for most languages. Thus, trained models perform poorly on the majority of the supported languages. Our work addresses this challenge by introducing a scalable pipeline for constructing speech datasets from parliamentary recordings. The proposed pipeline includes robust components for media retrieval and a two-stage alignment algorithm designed to handle non-verbatim transcripts and long-form audio. Applying this pipeline to recordings from 22 European parliaments, we extract over 61k hours of aligned speech segments, achieving substantial per-language coverage with 19 languages exceeding 1k hours and 22 languages exceeding 500 hours of high-quality speech data. We obtain an average 41.8% reduction in word error rates over baselines when finetuning an existing ASR model on our dataset, demonstrating the usefulness of our approach.

1 Introduction

Multilingual speech datasets have been essential for the recent progress in automatic speech recognition (ASR) and text-to-speech (TTS) model performance gains. The most significant advancements in ASR and TTS rely on large-scale training data, which is available for only a handful of high-resource languages. For the vast majority of languages, the lack of transcribed speech data poses a major obstacle to building reliable models. Recent large-scale ASR work [25], suggests that at least 1k hours of transcribed speech per language is typically required for modern ASR systems to reach acceptable performance. Table 1 compares several major multilingual speech datasets, illustrating a persistent imbalance. While some datasets span over 100 languages, only a small subset provide more than 1k hours of data for more than a few languages. For example, Common Voice [2] includes over 130 languages but only 8 of them exceed 1k hours. VoxPopuli [31], a benchmark dataset from parliamentary speech, includes 16 languages but none meet this threshold. This imbalance limits the ability of multilingual models to generalize well beyond a small set of dominant languages. Parliamentary proceedings present a compelling source of multilingual speech. Many governments make recordings and transcripts of their sessions publicly available, offering long-form speech in their official language. Moreover, most national parliaments provide more than 1k hours of data. However, building usable speech datasets from these sources is complex: The data is typically fragmented across platforms and formats, transcripts are often non-verbatim, and recordings are long and unsegmented. Current pipelines for dataset construction are limited to clean transcripts, which makes it difficult to scale across many parliaments, as individual transcript cleaning would be required.

In this work, we address these limitations through two primary contributions. First, we present a scalable, open-source pipeline for constructing speech datasets from parliamentary proceedings.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

Table 1: Comparison of Major Multilingual Speech Datasets. ">1k hrs" and ">500 hrs" columns indicate the number of languages exceeding these data volume thresholds. While some datasets contain more languages, EUROSPEECH provides the most languages above the respective data volume thresholds. The amount of languages over 500 and 1k hrs for the USM dataset is unknown.

Dataset	# Languages	Total Hours	>1k hrs	>500 hrs	Availability
MSR-86k [16]	15	86.3k	14	15	Public
Common Voice [2]	133	22.1k	8	15	Public
MLS [23]	8	50.0k	4	6	Public
FLEURS [4]	102	1.4k	0	0	Public
YODAS [17]	149	369.5k	13	15	Public
Emilia [10]	6	101.0k	5	6	Public
VoxPopuli [31]	16	1.8k	0	0	Public
GigaSpeech 2 [32]	3	30.0k	3	3	Public
Whisper Data [25]	91	680.0k	16	21	Private
BABEL [6]	$\sim$26	$\sim$1.0k	0	0	Private
USM [34]	73	90.0k	n/a	n/a	Private
MMS-Lab [24]	1107	44.7k	0	0	Private
EuroSpeech	**22**	**61k**	**19**	**22**	**Public**

The pipeline includes tools for downloading media and transcript files from diverse sources, as well as a robust alignment module featuring a two-stage dynamic alignment algorithm. The system supports various transcript formats (e.g., PDF, DOCX, SRT) via built-in, extensible parsers, and performs transcript normalization alongside optional LLM-based cleaning. It is specifically designed to be robust against non-verbatim transcripts and easily adaptable to new data sources with minimal engineering effort. Second, we introduce EUROSPEECH,[1] a new multilingual speech dataset built using our pipeline. EUROSPEECH comprises approximately 61k hours of aligned speech across 22 languages. Based on filtering by character error rate (CER), we obtain high-quality subsets including 50.5k hours at CER < 20%. This subset provides over 1k hours of data for 19 languages and over 500 hours for 22 languages. Together, these contributions provide a versatile pipeline for multilingual dataset creation and a valuable new resource for training and evaluating ASR and TTS models across a wider range of languages.

2 Related Work

2.1 Multilingual Speech Datasets

Initial ASR research featured single-language datasets such as Switchboard [9] and LibriSpeech [22], before progressing to multilingual efforts such as the IARPA BABEL program for low-resource languages [6]. Public multilingual corpora expanded with audiobook-derived MLS [23], the broadly crowdsourced Common Voice [2], and the benchmarking-focused FLEURS dataset [4]. Web-sourced data further increased scale in datasets such as YODAS [17], MSR-86k [16], GigaSpeech 2 [32], and Emilia [10], the last emphasizing diverse spontaneous speech.

Despite these advancements and large total reported hours, a critical imbalance persists: the large majority of languages in publicly available datasets contain below 1k hours of data per language, as detailed in Table 1. This skewness limits the development of high-performing multilingual ASR systems across many languages. The significant scale of private datasets (e.g., for Whisper [25], Google's USM [34], Meta's MMS-Lab [24]) highlights the benefits of extensive data but their inaccessibility underscores the need for large, open datasets.

Parliamentary speech, used in datasets such as VoxPopuli [31], Europarl-ASR [7], and various national corpora [14, 15, 30, 26, 11, 27, 8, 18], offers a promising domain-specific source of multilingual data. However, VoxPopuli, the largest publicly available multilingual parliamentary dataset only contains 1.8k hours across 16 languages, none of them exceeding the 1k hours threshold recommended for robust ASR training.

[1] EUROSPEECH is available at `https://huggingface.co/datasets/disco-eth/EuroSpeech`

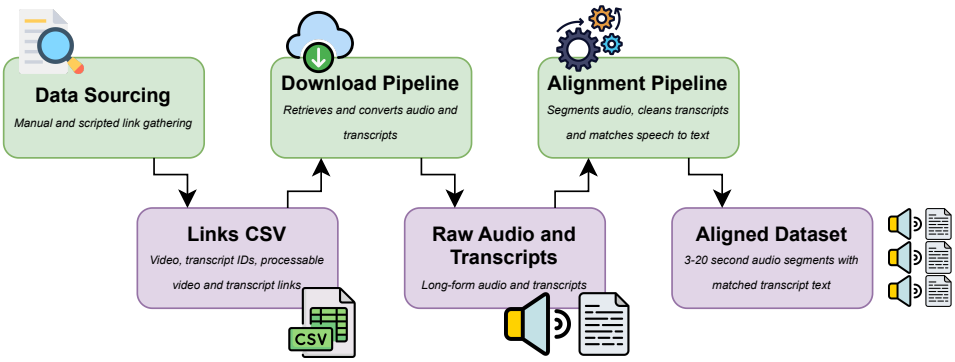

Figure 1: Overview of the EUROSPEECH data processing pipeline. The workflow begins with the initial **Data Sourcing and Metadata Collection** phase, which gathers metadata from parliamentary websites and APIs. This structured information (as **Links CSV**) feeds into the **Download Pipeline** to retrieve raw audio and transcripts. The **Raw Audio and Transcripts** are then processed by the **Alignment Pipeline**, which segments the audio and matches it to the corresponding text. The final output is the **Aligned Dataset**, consisting of short audio segments paired with their transcriptions, ready for model training.

EUROSPEECH differs from VoxPopuli in two key aspects. First, the scale and language coverage differ substantially: while VoxPopuli contains 1.8k hours with no languages exceeding 1k hours of transcribed data, EUROSPEECH provides 50.5k hours (CER < 20%) with 19 languages exceeding the 1k hour threshold. Second, the data sources differ fundamentally. VoxPopuli collected data from the European Union Parliament in Brussels, where representatives of all EU nations give speeches in their various languages, with recordings and transcripts sourced from a single website. In contrast, EUROSPEECH gathers parliamentary speeches from each country's individual national parliament, requiring custom scripts to collect data from 22 separate parliamentary websites.

Our work directly addresses these gaps by significantly increasing both the scale and language coverage of publicly available parliamentary speech data, as demonstrated by our EUROSPEECH dataset (Table 1).

2.2 Speech-Text Alignment Techniques

The construction of speech datasets requires precise alignment between raw audio and textual transcripts, which becomes particularly complex when processing noisy or non-verbatim sources such as parliamentary recordings. Alignment pipelines typically follow several stages: initial segmentation (via voice activity detection, speaker diarization, or fixed-duration chunking), matching audio segments to transcript sections (often through approximate ASR-based text matching), forced alignment (FA) for precise time-stamping, and finally, refined segmentation and quality filtering [23, 3, 31, 33, 24, 22, 7]. Some recent datasets (e.g., LibriHeavy [13]) used off-the-shelf ASR models such as Whisper for ASR-based audio-text matching. Notably, bootstrapping methods that use preliminary alignments to train better acoustic models for subsequent dataset re-alignment have proven effective [3, 24].

Our alignment pipeline incorporates a two-stage dynamic alignment approach that effectively handles noisy, non-verbatim transcripts typical of parliamentary speech, requiring minimal manual intervention. The approach is inspired by the dynamic CER-based matching employed by the VAC pipeline [33], yet extends it to handle the non-verbatim and noisy parliamentary transcripts robustly.

3 Data Collection Process

Building high-quality multilingual speech datasets from real-world parliamentary data introduces unique challenges: audio and transcripts are often long, noisy, and poorly aligned; metadata is

inconsistent or absent; and content is served through a wide range of web technologies. To address these issues, we design a scalable, modular system that transforms raw parliamentary recordings into clean, aligned speech-text segments suitable for model training. Our design prioritizes extensibility, fault isolation, and low operational overhead, enabling non-expert teams to replicate this process for new languages and data sources.

3.1 Pipeline Overview

Our proposed pipeline, shown in Figure 1, comprises an initial data sourcing and metadata collection phase, followed by two core automated pipelines: The **Download Pipeline** retrieves, standardizes, and converts raw audio and transcripts using a format-agnostic architecture. The **Alignment Pipeline** segments, transcribes, and aligns multi-hour recordings with non-verbatim transcripts using a noise-tolerant, dynamic matching algorithm.

These stages automatically process parliamentary data formats with minimal manual intervention. The two-stage alignment algorithm (details in Section 3.4) robustly handles heterogeneous data sources and non-verbatim transcripts, requiring only initial data source links as input. Once these links are collected, the pipeline performs all downloading, processing, and alignment steps automatically, eliminating the need for format-specific customization or manual transcript pre-processing.

3.2 Data Sourcing and Metadata Collection

Parliamentary data is published in a wide range of formats with little standardization across countries. Some parliaments provide direct access to downloadable MP4 files and clean transcripts in HTML; others publish only streaming video behind dynamic players or offer transcripts as scanned PDFs or DOCX documents. The goal of this initial data sourcing stage is to overcome the challenge of these diverse and unstandardized parliamentary sources. This is achieved by extracting and organizing key metadata (e.g., media and transcript URLs) into a standardized CSV file. This CSV then provides a consistent input format, enabling the subsequent automated download and alignment pipelines to operate uniformly, irrespective of the origin of the data.

The process begins with manual inspection of each parliament website, these can range from dedicated media portals to various scattered web pages. We identify data formats, access methods, and the structure of session-related information. This investigation, which can be challenging due to inconsistencies in how data is published and linked, informs the development of custom data collection scripts. These scripts aim to extract metadata for each session, which typically includes an audio or video URL, links to one or more potential transcripts, and a unique session identifier. We store this information in a standardized CSV format. The resulting CSV serves as the interface between this initial collection phase and the subsequent phases of the pipeline. By encapsulating all source-specific access logic and discovered links into this metadata file, the download and alignment pipelines can operate uniformly, allowing for a simple and efficient way of scaling to new parliaments.

3.3 Download Pipeline

Given the structured CSVs produced by the initial data sourcing and metadata collection stage, the download pipeline automates the retrieval of referenced audio, video, and transcript files. This stage is designed for robustness and extensibility across various types of content. The pipeline ingests diverse source types through a dispatch architecture, mapping URLs to specialized handlers. Built-in handlers cover common sources (e.g., direct links, YouTube, dynamic pages). Custom extractors handle parliamentary websites that require custom link extraction or transcript processing (e.g., special video players, multi-page transcript collection) without the need to change the core logic. Additionally, we implemented checkpointing, error recovery, and parallelization, as well as session-level download status, tracked in a central PostgreSQL database. These features enable distributed, non-redundant execution and automatic retries for failed downloads.

3.4 Alignment Pipeline

The alignment pipeline transforms raw audio and transcripts into short paired segments suitable for ASR and TTS model training. This stage addresses the challenges of long, noisy audio and diverse, non-verbatim transcripts as well as ambiguous mappings to create high-quality training data. Given

a long input recording (often between 1–10 hours) and a set of candidate transcripts (e.g., in PDF, HTML, or DOCX format), the pipeline first segments the audio into 3–20 second utterances using voice activity detection [28]. Each segment is then transcribed using an ASR model. We use Whisper Turbo [25] as the default ASR model, in the case of Maltese we use a fine-tuned Whisper Model [12]. These generated ASR text snippets are used to align the audio segments to the human transcript. Additionally, our proposed pipeline supports any ASR model as well as speaker diarization, which is useful to ensure single speaker audio segments. Raw transcripts are automatically preprocessed into cleaned and standardized text. The pipeline includes built-in parsers for common formats (e.g., PDF, DOCX, HTML, TXT, SRT), with optional LLM-based cleaning to remove non-speech elements from documents such as PDFs (cf. Appendix D). This stage of the pipeline is easily extensible for new formats or custom logic.

We propose a **two-stage dynamic algorithm** to align ASR generated transcripts with parliamentary transcripts (detailed algorithm pseudo-code in Appendix B). The parliamentary transcripts often contain speech content mixed with non-verbatim text, unuttered text, or speaker annotations. This algorithm enables data collection from diverse sources by eliminating the need for manual transcript pre-processing. Unlike VAC [33], which uses a single dynamic window for alignment, our approach uses a two-stage coarse-to-fine strategy to identify matching transcript text for each audio segment:

1. **Coarse Search:** Uses a sliding window of size n, where n is the length of the current ASR generated transcript. The starting position for the sliding window is set to the last matched position of the previous segment. This coarse search identifies candidate text spans by computing the CER between each transcript window and the ASR generated transcript, skipping transcript sections with high error rates. We pass the first candidate that has a CER below 30% to the refined search stage. If no candidates below 30% CER are found, we pass the k (default k = 3) candidates with the lowest CER to the refined stage.
2. **Refined Search**: The coarse candidate windows identified in the first stage are adjusted to find the lowest possible CER within the respective window. For each candidate window, we test different start positions and window sizes within a local margin. Specifically, the start position is varied within ± 15 words around the coarse window's starting position, and for each start position, we test window sizes ranging from $(L - 15)$ to $(L + 15)$ words, where L represents the ASR segment length in words. We select the start position and window size that minimizes CER between the transcript span and ASR generated transcript.

A fallback mechanism restarts this two-stage search for the current segment. The key difference is that the starting position for the coarse search is set to the beginning of the entire transcript rather than the last matched position from the previous segment. The fallback mechanism addresses cases where previous misalignments positioned the starting position beyond the current segment's true location in the transcript. Restarting from the beginning allows the algorithm to find the correct match instead of searching past it.

Our alignment algorithm requires pairing each audio file with a single transcript file. However, parliaments may provide multiple transcript files for a given day, and it is often not clear which of these transcripts belongs to which audio file. For example, when transcripts and audio files can only be matched by their publication date and multiple files share the same date, we cannot clearly determine which transcript belongs to which audio. Additionally, individual transcripts are often available in multiple formats (e.g., PDF, HTML, DOCX). To automate finding the correct pairings, we use a two-stage selection process: we first align the audio to all candidate transcripts in all available formats, then select the best format for each transcript based on median CER, and finally choose the transcript to use based on configurable criteria (e.g., lowest CER or all transcripts below a CER threshold).

The output of this alignment stage includes a JSON summary for each audio file and detailed alignment files for each selected transcript. Each file contains timestamps, matched text, and quality metrics. These files form the basis for downstream filtering and dataset generation.

3.5 Filtering and Output Format

Once segment-level alignments have been established, the final step is to filter and package the aligned data for training. Although the alignment algorithm produces timestamps and CER estimates for each

Table 2: Overview of EUROSPEECH: Aligned audio hours per language at our three quality character error rate (CER) Thresholds. Languages are sorted by hours at CER < 20% which represents the main subset of EUROSPEECH.

Language	Code	Total Aligned (h)	CER < 30% (h)	CER < 20% (h)	CER < 10% (h)
Croatia	hr	7484.9	5899.7	5615.8	4592.0
Denmark	da	7014.2	6435.0	5559.8	3443.7
Norway	no	5326.2	4578.8	3866.7	2252.2
Portugal	pt	5096.3	4036.7	3293.5	2105.9
Italy	it	4812.8	3539.6	2813.7	1767.3
Lithuania	lt	5537.9	3971.0	2681.2	956.6
United Kingdom	en	5212.2	3790.7	2609.3	1175.0
Slovakia	sk	2863.4	2722.4	2553.6	2070.8
Greece	el	3096.7	2717.6	2395.4	1620.9
Sweden	sv	3819.4	2862.6	2312.8	1360.1
France	fr	5476.8	2972.1	2249.8	1347.6
Bulgaria	bg	3419.6	2570.4	2200.1	1472.8
Germany	de	2472.2	2354.2	2184.4	1698.4
Serbia	sr	2263.1	1985.1	1855.7	1374.1
Finland	fi	2130.6	1991.4	1848.2	1442.2
Latvia	lv	2047.4	1627.9	1218.8	499.9
Ukraine	uk	1287.8	1238.3	1191.1	1029.8
Slovenia	sl	1338.2	1241.7	1156.4	900.5
Estonia	et	1701.1	1430.9	1014.9	382.5
Bosnia & Herz.	bs	860.2	781.9	691.3	447.8
Iceland	is	1586.1	974.1	647.4	171.4
Malta	mt	3281.6	1284.3	613.0	143.9
Total		**78128.6**	**61006.4**	**50572.9**	**32255.5**

utterance, not all matched pairs are equally reliable, especially given transcript noise, overlapping speech, or ASR errors.

We adopt CER-based filtering as the primary mechanism for quality control. A threshold (e.g., CER $< 20\%$) is applied for each segment, allowing users to select a desired trade-off between dataset size and quality. Additionally, we track total aligned duration at multiple quality tiers (e.g., CER $< 10\%$, $< 20\%$, $< 30\%$), enabling granular evaluation and filtering without rerunning the alignment pipeline.

For each audio file we output a summary JSON file containing overall alignment statistics and references to all candidate transcripts. For each accepted audio-transcript pair, we generate a separate alignment JSON file listing all aligned segments with their start and end times, ASR text, matched human transcript text, and CER.

4 The EUROSPEECH Dataset

Running our proposed data processing pipeline described in Section 3, we constructed EUROSPEECH, a new multilingual speech corpus derived from parliamentary proceedings from 22 European nations. To create EUROSPEECH we processed as many publicly available parliamentary recordings as possible. After the initial alignment and segmentation process, which removed silences and unaligned portions, we obtained approximately 78k hours of aligned speech-text data. Finally, we then created quality-filtered subsets based on Character Error Rate (CER):

- CER < 30%: approximately 61k hours (78.2% of aligned data)
- CER < 20%: approximately 51k hours (65.4% of aligned data)
- CER < 10%: approximately 32k hours (41.0% of aligned data)

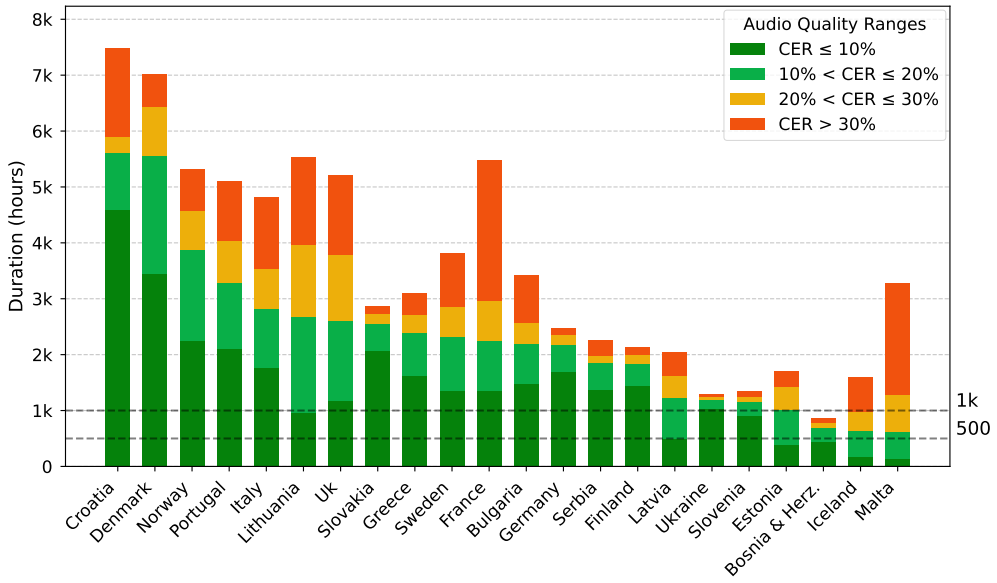

Figure 3: Audio duration for each language in EUROSPEECH after the alignment pipeline and at different Character Error Rate (CER) filtering stages (CER < 30%, < 20%, and < 10%). Languages are ordered by their data volume at CER < 20%. A dashed horizontal lines indicates the 1,000-hour and 500-hour thresholds. EUROSPEECH showcases large amounts of low-CER data across its languages.

The CER < 20% subset serves as our primary dataset for comparisons, we chose the 20% threshold based on VoxPopuli [31]. Within this subset, EUROSPEECH provides >1k hours of data for 19 languages and >500 hours for 22 languages. Table 2 presents a detailed breakdown of the dataset composition for each language.

The audio segments in EUROSPEECH typically range from 3 to 20 seconds, durations suitable for training ASR and TTS models. The audio in the published dataset is sampled at 16 kHz, which is the standard sampling rate for training ASR models. We plan to upload a 24 kHz version of the dataset as this is most common for TTS models. The data reflects the formal speaking style characteristic of parliamentary debates. The European coverage of the data is shown in Figure 2.

Unlike many existing multilingual datasets that are heavily skewed toward a few high-resource languages, EUROSPEECH maintains a more equitable distribution across languages, a key differentiator of our corpus. Figure 3 further details the data quantities per language and the impact of CER filtering stages per language.

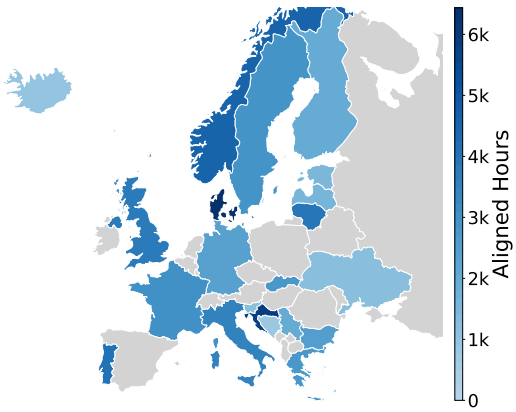

Figure 2: The broad European geographic coverage of EUROSPEECH. Countries are colored by the total hours of aligned speech data (CER < 30%) available in the dataset. A perceptually uniform color scale is used.

To facilitate standardized benchmarking, we provide predefined train, development, and test splits for each language. To ensure data integrity and prevent leakage between sets, these splits are constructed by assigning entire parliamentary sessions (i.e., all segments derived from a single original long audio recording) exclusively to one of the train, development, or test sets. The proportions for these splits follow common practices and are detailed in the dataset repository.

Table 3: Word Error Rates (%) of Whisper v3 Turbo on the FLEURS test set before and after finetuning on EUROSPEECH. Finetuned models consistently reduce WER across all evaluated languages, demonstrating the practical value of the dataset for improving multilingual ASR performance.

Language	Baseline	Finetuned	Rel. Improvement
Maltese	72.2	25.9	64.1%
Icelandic	20.0	15.0	25.0%
Lithuanian	25.0	15.9	36.4%
Latvian	19.3	11.1	42.5%
Slovenian	20.5	13.0	36.7%
Estonian	18.4	9.9	46.1%
Average	29.2	15.1	41.8%

5 Finetuning ASR Models with EUROSPEECH

This section evaluates the effectiveness of our EUROSPEECH dataset for improving ASR performance, particularly on low-resource European languages. We demonstrate that finetuning pretrained multilingual ASR model Whisper v3 Turbo on the EUROSPEECH dataset yields considerable improvements in transcription performance.

5.1 Experimental Setup

We evaluate the impact of our dataset by finetuning the pretrained Whisper v3 Turbo model [25] on six European languages from our collection: Maltese, Icelandic, Lithuanian, Latvian, Slovenian, and Estonian. These languages were selected to represent different language families and focus on low-resource availability levels in the existing literature.

Baseline Model: We use Whisper v3 Turbo as our baseline for finetuning. This model was pretrained on 680k hours of labeled audio in 98 languages but has limited exposure to many European languages in our dataset. As of the time of writing, performance on the FLEURS [4] test set was not publicly reported for certain languages targeted in our finetuning. Consequently, we conducted independent baseline evaluations of the Whisper v3 Turbo model on FLEURS.

Finetuning: For each of the six languages we evaluated, approximately 200 hours of training data with the lowest CER were selected for training. More details can be found in Appendix A.

Evaluation: We evaluate the models on the FLEURS test set. For evaluation metrics, we report the Word Error Rate (WER). To ensure fair comparison across languages with different writing systems and word formation patterns, we apply the NFKC normalization process as used in the training of all Whisper models before computing error rates.

5.2 Results

Table 3 presents the performance of the finetuned Whisper v3 Turbo compared to the baseline on the FLEURS test set. We observe that, finetuning on EUROSPEECH substantially improves transcription performance across all tested languages. On average, we observe a relative WER reduction of 41.8% on the out-of-domain FLEURS test set. The results demonstrate that our dataset enables significant WER improvements, achieving competitive performance for open-source ASR models, while using a limited subset of EUROSPEECH's training data.

6 Limitations

While EUROSPEECH represents a balanced dataset containing various low-resource languages, certain limitations remain. The dataset is derived entirely from parliamentary recordings, a domain characterized by formal, planned, and often repetitive speech. This linguistic register, while useful for certain modeling tasks, may not adequately represent the diversity of natural spoken language encountered in more spontaneous or informal settings. As such, models trained solely on EuroSpeech may exhibit reduced performance when deployed in conversational or non-scripted speech scenarios.

The dataset's linguistic and geographic scope, although broad within the European context, remains limited in global coverage. Many underrepresented languages, particularly those outside of Europe, are not included. Even within the covered languages, variation in dialect, sociolect, and regional accents is likely constrained by the nature of parliamentary speech, which tends to reflect standard or official varieties. This may affect the robustness and fairness of models trained on the dataset, particularly in settings requiring sensitivity to linguistic diversity.

From a technical perspective, the alignment process is dependent on the quality of existing ASR models, which are used to generate intermediate transcriptions that are needed to align the human transcript with the correct audio segment. While the proposed alignment algorithm is designed to be robust to non-verbatim transcripts and noisy inputs, its performance is ultimately constrained by the capabilities of the underlying ASR models, which can vary significantly across languages and acoustic conditions. In low-resource languages or in instances of degraded audio quality, the accuracy of alignments may be reduced, potentially impacting the quality of the resulting training data.

7 Conclusions

In this work, we presented a source-agnostic, open-source pipeline for speech-text alignment that can process any audio with potentially matching transcripts. Its robust two-stage alignment algorithm and modular architecture enables non-experts to create high-quality speech datasets for diverse applications and languages. Using our proposed pipeline, we created EUROSPEECH, a multilingual speech corpus containing over 50.5k hours of aligned parliamentary speech with CER < 20% across 22 European languages. Unlike existing public datasets, EUROSPEECH provides substantial data for all included languages, with 19 languages exceeding 1k hours and 22 exceeding 500 hours. The balanced distribution across languages addresses the severe imbalance in current multilingual speech resources. To demonstrate the usefulness of our dataset, we finetune an ASR model on EUROSPEECH for six low-resource languages, showing an average 41.8% reduction in word error rates over baselines. The pipeline codebase as well as the dataset are made publicly available.

We believe that both the EUROSPEECH dataset and our proposed pipeline can be useful starting points for further work on multilingual and low-resource speech processing. In the future, the pipeline could be extended to other domains such as conversational speech, or adapted to include more languages and metadata such as speaker or session information. Making these tools available can help lower the barrier to building high-quality datasets, especially for underrepresented languages.

References

[1] Tanel Alumäe, Joonas Kalda, Külliki Bode, and Martin Kaitsa. Automatic closed captioning for Estonian live broadcasts. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 492–499, Tórshavn, Faroe Islands, May 2023. University of Tartu Library. URL https://aclanthology.org/2023.nodalida-1.49.

[2] Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In *LREC*, 2020. URL https://arxiv.org/abs/1912.06670.

[3] Guoguo Chen, Shuzhou Chai, Guanbo 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, Yujun Wang, Zhao You, and Zhiyong Yan. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio, 2021. URL https://arxiv.org/abs/2106.06909.

[4] Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. In *EMNLP*, 2022. URL https://arxiv.org/abs/2205.12446.

[5] Marco Gaido, Sara Papi, Luisa Bentivogli, Alessio Brutti, Mauro Cettolo, Roberto Gretter, Marco Matassoni, Mohamed Nabih, and Matteo Negri. Mosel: 950,000 hours of speech data for open-source speech foundation model training on eu languages. *arXiv preprint arXiv:2410.01036*, 2024.

[6] M. J. F. Gales, K. M. Knill, A. Ragni, and S. P. Rath. Speech recognition and keyword spotting for low-resource languages : Babel project research at cued, May 2014. URL https://www.isca-speech.org/archive/sltu%5f2014/sl14%5f016.html. © 2014 ISCA. Reproduced in accordance with the publisher's self-archiving policy.

[7] Gonçal V. Garcés Díaz-Munío, Joan-Albert Silvestre-Cerdà, Javier Jorge, Adrià Giménez Pastor, Javier Iranzo-Sánchez, Pau Baquero-Arnal, Nahuel Roselló, Alejandro Pérez-González-de Martos, Jorge Civera, Albert Sanchis, and Alfons Juan. Europarl-asr: A large corpus of parliamentary debates for streaming asr benchmarking and speech data filtering/verbatimization. *Interspeech*, pages 3695–3699, 2021. doi: 10.21437/Interspeech.2021-1905.

[8] Diana Geneva, Georgi Shopov, and Stoyan Mihov. Building an asr corpus based on bulgarian parliament speeches. In *International Conference on Statistical Language and Speech Processing*, 2019. URL https://api.semanticscholar.org/CorpusID:203565884.

[9] John J. Godfrey, Edward C. Holliman, and Jane McDaniel. Switchboard: telephone speech corpus for research and development. *Proceedings of the 1992 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 517–520, 1992.

[10] Haorui He, Zengqiang Shang, Chaoren Wang, Xuyuan Li, Yicheng Gu, Hua Hua, Liwei Liu, Chen Yang, Jiaqi Li, Peiyang Shi, Yuancheng Wang, Kai Chen, Pengyuan Zhang, and Zhizheng Wu. Emilia: An extensive, multilingual, and diverse speech dataset for large-scale speech generation. In *SLT*, 2024. URL https://arxiv.org/abs/2407.05361.

[11] Inga Rún Helgadóttir, Róbert Kjaran, Anna Björk Nikulásdóttir, and Jón Guðnason. Building an asr corpus using althingi's parliamentary speeches. *Interspeech*, pages 2163–2167, 2017. doi: 10.21437/Interspeech.2017-903.

[12] Carlos Daniel Hernandez Mena. Acoustic model in maltese: whisper-largev2-maltese-8k-steps-64h., 2023. URL https://huggingface.co/carlosdanielhernandezmena/whisper-largev2-maltese-8k-steps-64h.

[13] Wei Kang et al. Libriheavy: A 50,000 hours asr corpus with punctuation casing and context. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10991–10995. IEEE, 2024.

[14] Andreas Kirkedal, Marija Stepanović, and Barbara Plank. Ft speech: Danish parliament speech corpus. *Interspeech*, pages 442–446, 2020. doi: 10.21437/Interspeech.2020-3164.

[15] Baybars Kulebi, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. Parlamentparla: A speech corpus of catalan parliamentary sessions. *Proceedings of the Workshop ParlaCLARIN III*, pages 125–130, 2022. URL https://aclanthology.org/2022.parlaclarin-1.18/.

[16] Song Li, Yongbin You, Xuezhi Wang, Zhengkun Tian, Ke Ding, and Guanglu Wan. Msr-86k: An evolving, multilingual corpus with 86,300 hours of transcribed audio for speech recognition research. In *Interspeech*, 2024. URL https://arxiv.org/abs/2406.18301.

[17] Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. Yodas: Youtube-oriented dataset for audio and speech. In *ASRU*, 2023. URL https://arxiv.org/abs/2406.00899.

[18] Nikola Ljubešić, Peter Rupnik, and Danijel Koržinek. The parlaspeech collection of automatically generated speech and text datasets from parliamentary proceedings. *Speech and Computer*, pages 137–150, 2024. doi: 10.1007/978-3-031-77961-9_10.

[19] Carlos Mena, Albert Gatt, Andrea DeMarco, Claudia Borg, Lonneke Van der Plas, Amanda Muscat, and Ian Padovani. Masri-headset: A maltese corpus for speech recognition. *arXiv preprint arXiv:2008.05760*, 2020.

[20] Carlos Daniel Hernández Mena, Þorsteinn Daði Gunnarsson, and Jón Guðnason. Samrómur milljón: An asr corpus of one million verified read prompts in icelandic. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14305–14312, 2024.

[21] National Library of Norway. Stortinget speech corpus version 1.0, 2023. URL https://www.nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-91/. Norwegian Language Bank.

[22] Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: An asr corpus based on public domain audio books. *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5206–5210, 2015. URL https://api.semanticscholar.org/CorpusID:2191379.

[23] Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411, 2020. URL https://api.semanticscholar.org/CorpusID:226202134.

[24] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *arXiv preprint arXiv:2305.13516*, 2023. URL https://arxiv.org/abs/2305.13516.

[25] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*, 2022. URL https://arxiv.org/abs/2212.04356.

[26] Faton Rekathati. The kblab blog: Rixvox: A swedish speech corpus with 5500 hours of speech from parliamentary debates, 2023. URL https://kb-labb.github.io/posts/2023-03-09-rixvox-a-swedish-speech-corpus/.

[27] Per Erik Solberg and Pablo Ortiz. The Norwegian parliamentary speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1003–1008, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.106/.

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

[29] 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. ASR database ARTUR 1.0 (transcriptions), 2023. ISSN 2820-4042. URL http://hdl.handle.net/11356/1772. Slovenian language resource repository CLARIN.SI.

[30] Anja Virkkunen, Aku Rouhe, Nhan Phan, and Mikko Kurimo. Finnish parliament asr corpus. *Language Resources and Evaluation*, pages 1–26, 2022. URL https://api.semanticscholar.org/CorpusID:247779318.

[31] Changhan Wang, Morgane Rivière, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *ACL*, 2021. URL https://arxiv.org/abs/2101.00390.

[32] Yifan Yang, Zheshu Song, Jianheng Zhuo, Mingyu Cui, Jinpeng Li, Bo Yang, Yexing Du, Ziyang Ma, Xunying Liu, Ziyuan Wang, et al. Gigaspeech 2: An evolving, large-scale and multi-domain asr corpus for low-resource languages with automated crawling, transcription and refinement. *arXiv preprint arXiv:2406.11546*, 2024.

[33] Ara Yeroyan and Nikolay Karpov. Enabling asr for low-resource languages: A comprehensive dataset creation approach, 2024. URL https://arxiv.org/abs/2406.01446.

[34] Yu Zhang et al. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*, 2023. URL https://arxiv.org/abs/2303.01037.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The main claim (a diverse dataset with sufficient data for low-resource languages) is supported by the collected dataset. We further validate the quality of the dataset by training ASR models on it.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The limitations of the work is discussed in Section 6.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not contain theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Details on the experimental setup are provided in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We make the complete dataset available on Huggingface. We make the code for our toolkit available on GitHub: `https://github.com/SamuelPfisterer/EuroSpeech`.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: The setup of experiments in section 5 was described to necessary detail in order to appreciate the results. The full details are provided in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: Statistical significance tests are not necessary to support our claim of improved performance for the finetuning.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The computer resources used for the experiments in section 5 and the data collection are reported in Appendix A.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The dataset introduced in the paper consists entirely of publicly available transcripts and audio material provided by the parliaments of the respective countries. Although our pipeline aims to minimize identifiable content, speakers may refer to individuals by name.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Broader impacts are discussed in Section 6 and Appendix E.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The dataset consists of official publicly available parliament audio and transcripts.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Data sources and licenses are discussed in Appendix C.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide detailed documentation of how to use our dataset and pipeline on Huggingface (`https://huggingface.co/datasets/disco-eth/EuroSpeech`) and GitHub.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: LLMs played a supporting role in writing the code necessary to create the dataset and pipeline introduced in the paper.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

A Experiment Setup

Data Collection Infrastructure. Download jobs were allocated 2 CPUs and 8GB RAM each. The total computational cost for data sourcing comprised approximately 4,200 hours across video downloads (3,930 hours) and transcript retrieval (280 hours). These estimates are based on job logging in our database and provide approximate resource requirements for replication.

Alignment Pipeline Compute. Audio-text alignment was performed using heterogeneous GPU resources including GeForce RTX 2080 Ti (11GB), Tesla V100 (32GB), Titan XP (12GB), Titan RTX (24GB), and RTX 3090 (24GB). The majority of computation utilized RTX 2080 Ti and Tesla V100 cards. Total alignment processing required approximately 5,548 GPU-hours across all jobs and languages.

ASR Model Fine-tuning. We fine-tuned Whisper v3 Turbo[2] using the following hyperparameters: batch size 64, gradient accumulation steps 2, learning rate 1e-5, warmup ratio 0.06, and linear learning rate scheduling. Training details for each language are provided in Table 4. All trainings were performed on NVIDIA RTX A6000 GPU cards.

Language selection was motivated by two factors: (1) these six languages exhibited the highest baseline WER with Whisper v3 Turbo, allowing demonstration of meaningful improvements with limited computational resources, and (2) poor baseline ASR performance creates additional challenges for our alignment pipeline, as ASR transcriptions for these languages contain more errors, providing a rigorous test of our pipeline's ability to match noisy ASR outputs to the correct segments in human transcripts.

Table 4: Fine-tuning configuration per language

Language	Training Data (h)	Epochs	Training Time (h)
Maltese	143	0.2	1.3
Icelandic	213	1.6	13.2
Lithuanian	365	2.5	43.2
Latvian	203	2.4	21.0
Slovenian	289	3.0	40.5
Estonian	262	2.8	28.6

[2] `https://huggingface.co/openai/whisper-large-v3-turbo`

B Data Collection Process

Algorithm 1 Two-Stage Dynamic Alignment Algorithm

Require: ASR segments S_{asr}, Full Transcript Text T
Require: CER threshold θ
Ensure: List of Aligned Segments $S_{aligned}$

$S_{aligned} \leftarrow \emptyset$
$last\_end\_idx \leftarrow 0$ ▷ End index of last matched transcript segment
for all segment $s_{asr} \in S_{asr}$ **do**
    ▷ Stage 1: Coarse Search (sequential from $last\_end\_idx$)
    $candidates \leftarrow$ CoarseSearch(s_{asr}, T, start_idx=$last\_end\_idx$)
    ▷ Stage 2: Refining within candidate regions
    $match \leftarrow$ RefinedSearch($candidates, s_{asr}, T$)
    if $match.cer > \theta$ **then** ▷ Fallback 1: Global Coarse Search
        $candidates_{global} \leftarrow$ CoarseSearch(s_{asr}, T, start_idx=0)
        $match_{global} \leftarrow$ RefinedSearch($candidates_{global}, s_{asr}, T$)
        if $match_{global}.cer > \theta$ **then** ▷ Global search also insufficient
            $match \leftarrow$ DefaultMatch($s_{asr}, T, last\_end\_idx$) ▷ Fallback 2
        else
            $match \leftarrow match_{global}$ ▷ Use global match
        end if
    end if
    Append $match$ to $S_{aligned}$
    $last\_end\_idx \leftarrow match.end\_idx$ ▷ Update for next sequential search
end for
return $S_{aligned}$

The two-stage dynamic alignment algorithm matches ASR-transcribed audio segments to corresponding segments in human transcripts. Processing segments sequentially, it maintains $last\_end\_idx$ to track transcript position and leverage temporal ordering. For each segment, coarse search employs a sliding window from the last matched position to identify candidate spans with minimal character error rate. Refined search then exhaustively searches over start position offsets and window lengths within a local margin around the candidate region to minimize character error rate. If the resulting alignment exceeds threshold θ a global search across the entire transcript is attempted. When quality thresholds cannot be met, default matching performs refined search around $last\_end\_idx$ and stores the best available match regardless of CER, ensuring complete dataset coverage.

C Data Sources

We sourced the parliamentary data primarily from the respective parliament websites of each country, with some additional content obtained from YouTube channels operated by the parliaments. For each country, we maintain a CSV file that lists all source links for video/audio files and transcript documents.

The video_id and transcript_id values present in the final EUROSPEECH dataset can be used to trace back to the specific source URLs for each audio segment and its corresponding text.

All CSV files containing the source metadata are publicly available on Hugging Face.[3] These files provide complete transparency regarding the origins of our dataset and enable others to replicate or extend our data collection methodology.

For copyright and licensing information regarding the parliamentary data from each country, we refer to Table 5 below, which details the relevant legal frameworks and licensing terms for each parliamentary source.

[3] `https://huggingface.co/datasets/SamuelPfisterer1/EuroSpeech-Data-Sources`

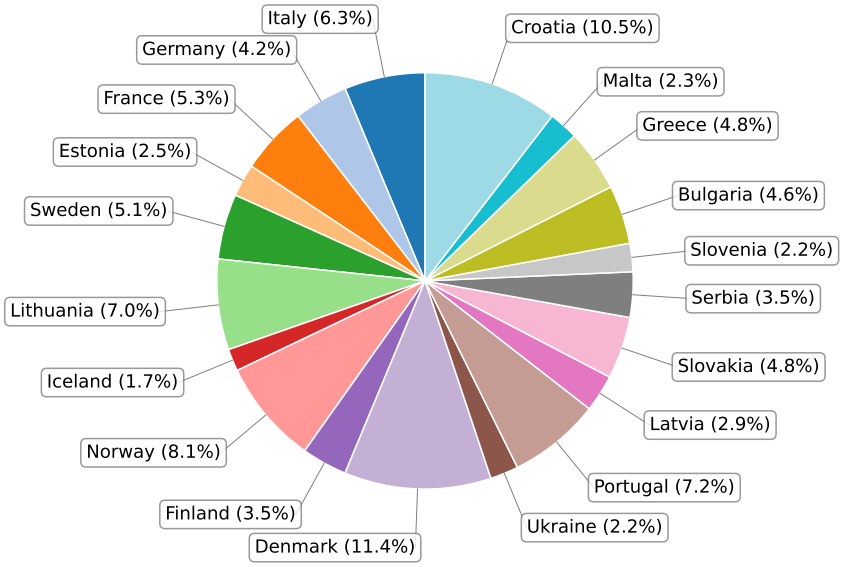

Figure 4: Language distribution in the EUROSPEECH CER < 30% subset, highlighting a key strength of our dataset: the balanced distribution across multiple languages rather than concentration in just a few dominant ones.

D LLM-Based Transcript Cleaning

Our pipeline provides an optional LLM-based cleaning feature specifically for PDF transcripts. When processing PDF transcripts, the pipeline first extracts text from each page. When LLM-based cleaning is used the extracted text is passed through a large language model with specific instructions to retain only spoken dialogue. We use Gemini Flash 2.0 as the default model for this cleaning step as we found it achieved a good performance-to-cost ratio.

The default system prompt used for LLM-based cleaning is:

> You are a multilingual assistant specialized in processing parliamentary transcripts. Your task is to clean the provided transcript page by removing all unnecessary metadata, annotations etc. while preserving only the literal spoken dialogue. Please follow these instructions: Remove the speaker labels that appear as headers before each speaker's dialogue. Remove all annotations, procedural notes, timestamps, and non-verbal cues. Ensure that only and all the spoken dialogue is in your response. Respond in the same language as the input and do not alter the spoken text.

The system prompt can be customized through the pipeline configuration. Tests based on one German parliamentary session showed median CER improvements from 12.3% to 9.7% for the final aligned segments when using LLM-based cleaning compared to standard PDF text extraction.

E Broader Impacts

This work aims to address the substantial imbalance in multilingual speech resources by introducing a large-scale, publicly available dataset with strong per-language coverage across 22 European languages. The EUROSPEECH corpus enables the development and evaluation of speech models for languages that have previously lacked sufficient training data, particularly in the context of automatic speech recognition (ASR) and text-to-speech (TTS) systems. By improving model performance

Table 5: Copyright of parliament data

Country	Source
Croatia	Legal Notice
Denmark	Legal Notice
Norway	NLOD License
Portugal	Portuguese Copyright Code Article 75
Italy	Italian Parliament Website references CC By 4.0 License
Lithuania	Republic of Lithuania Law on Copyright and Related Rights Article 22
United Kingdom	Terms and Conditions for audio, Open Government Licence for transcripts
Slovakia	Slovak Copyright Act Chapter One Section 5e)
Greece	Greek Copyright Law Article 2(5) and Article 25(1)(b)
Sweden	Law (2022:818)
France	License Ouverte
Bulgaria	Copyright Policy references CC BY 2.5 BG
Germany	Terms of Use
Serbia	Serbian Law on Copyright and Related Rights. Article 6(2)
Finland	Copyright Act Article 9, 22, and 25
Latvia	Latvian Copyright Law Section 21
Ukraine	Law of Ukraine on Copyright and Related Rights Article 8(1)(3)
Slovenia	Copyright and Related Rights Act Article 46-51
Estonia	Copyright Act, Estonian Youtube references CC BY SA
Bosnia & Herz.	Copyright Law Article 44 and 47
Iceland	Copyright Act Article 22
Malta	Re-Use of Public Sector Information Act Chapter 546

for under-resourced languages, the dataset has the potential to broaden access to speech technology and reduce the reliance on high-resource language data in multilingual systems. The dataset's origin in parliamentary recordings makes it well-suited for applications related to public sector accessibility, such as transcription and translation of government proceedings. However, this domain-specificity also imposes limitations: the speech style is formal, planned, and typically reflects standard language varieties. As a result, models trained exclusively on EUROSPEECH may generalize poorly to conversational or informal speech, and may underperform for dialectal, regional, or sociolectal variation not represented in parliamentary discourse.

The dataset is constructed from publicly available government media and does not include private or crowd-sourced content. Nevertheless, identifiable individuals may be mentioned in the transcripts, and downstream uses involving speaker identification or synthesis warrant careful consideration. While the primary goal is to support inclusive and transparent research, we acknowledge that speech models trained on EUROSPEECH could be used for purposes such as synthetic speech generation, which carries misuse potential in contexts such as impersonation or disinformation. The domain constraints of the data mitigate some of this risk, but further safeguards may be necessary depending on downstream applications.

Overall, this work contributes infrastructure that can lower the barrier to entry for multilingual speech research, particularly for low-resource languages. At the same time, it highlights the need for complementary datasets that capture greater linguistic diversity and less formal speech styles to support broader and more equitable generalization.

F Comparison with Existing Language-Specific Datasets

Table 6 presents a comprehensive comparison between EUROSPEECH and the largest publicly available speech datasets for each language in our corpus. This comparison demonstrates the substantial increase in available training data that EUROSPEECH provides for many European languages.

We created new state-of-the-art duration datasets for 12 languages, crossing the 1k hour threshold for 8 languages where previous datasets were below this threshold. For 5 languages, our durations are 10

Table 6: Comparison of EUROSPEECH with state-of-the-art speech datasets per language. Hours shown for EUROSPEECH correspond to the CER < 20% filtered subset. Bold values indicate cases where EUROSPEECH provides more data than existing datasets.

Country/Language	SOTA Dataset Name	SOTA Dataset Hours	EUROSPEECH Hours
Croatia	ParlaSpeech-HR [18]	3061	**5615.8**
Denmark	FT-Speech [14]	1800	**5559.8**
Norway	Stortinget Corpus [21]	5190	3866.7
Portugal	MOSEL [5]	5492	3293.5
Italy	MOSEL [5]	3756	2813.7
Lithuania	Common Voice [2]	25	**2681.2**
United Kingdom	MOSEL [5]	437238	2609.3
Slovakia	MOSEL [5]	61	**2553.6**
Greece	YODAS [17]	126.75	**2395.4**
Sweden	RixVox-v2 [26]	22900	2312.8
France	MOSEL [5]	26984	2249.8
Bulgaria	BG-PARLAMA [8]	249	**2200.1**
Germany	MOSEL [5]	9236	2184.4
Serbia	ParlaSpeech-RS [18]	896.22	**1855.7**
Finland	Finnish Parliament ASR [30]	3087	1848.2
Latvia	Common Voice [2]	263	**1218.8**
Ukraine	YODAS [17]	396.598	**1191.1**
Slovenia	ASR database ARTUR 1.0 [29]	884	**1156.4**
Estonia	TalTech Speech Dataset [1]	1334	1014.9
Bosnia & Herz.	YODAS [17]	9.37	**691.3**
Iceland	Samrómur Milljón [20]	967	647.4
Malta	MASRI [19]	44	**613.0**

to 100 times greater than those of prior state-of-the-art datasets. It is also important to highlight that some of the previous state-of-the-art datasets are not as easily usable (i.e., they do not have a unified representation and cannot be used with a few lines of code from HuggingFace). Furthermore, the quality on some of these datasets is difficult to verify as they do not explain how they filtered the dataset. The EUROSPEECH durations in this table refer to the 20% CER subset which is equivalent to the threshold used for VoxPopuli [31].