

722 A Experiment Setup

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

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

732 **ASR Model Fine-tuning.** We fine-tuned Whisper v3 Turbo² using the following hyperparameters:
733 batch size 64, gradient accumulation steps 2, learning rate 1e-5, warmup ratio 0.06, and linear
734 learning rate scheduling. Training details for each language are provided in Table 4. All trainings
735 were performed on NVIDIA RTX A6000 GPU cards.

736 Language selection was motivated by two factors: (1) these six languages exhibited the highest
737 baseline WER with Whisper v3 Turbo, allowing demonstration of meaningful improvements with
738 limited computational resources, and (2) poor baseline ASR performance creates additional challenges
739 for our alignment pipeline, as ASR pseudo-labels for these languages contain more errors, providing
740 a rigorous test of alignment robustness.

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>

741 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 $S_{aligned}$

```

1:  $S_{aligned} \leftarrow \emptyset$ 
2:  $last\_end\_idx \leftarrow 0$  ▷ End index of last matched transcript segment
3: for all segment  $s_{asr} \in S_{asr}$  do
4:   ▷ Stage 1: Coarse Search (sequential from  $last\_end\_idx$ )
5:    $candidates \leftarrow \text{CoarseSearch}(s_{asr}, T, \text{start\_idx}=last\_end\_idx)$ 
6:   ▷ Stage 2: Fine-Tuning within candidate regions
7:    $match \leftarrow \text{FineTune}(candidates, s_{asr}, T)$ 
8:   if  $match.cer > \theta$  then ▷ Fallback 1: Global Coarse Search
9:      $candidates_{global} \leftarrow \text{CoarseSearch}(s_{asr}, T, \text{start\_idx}=0)$ 
10:     $match_{global} \leftarrow \text{FineTune}(candidates_{global}, s_{asr}, T)$ 
11:    if  $match_{global}.cer > \theta$  then ▷ Global search also insufficient
12:       $match \leftarrow \text{DefaultMatch}(s_{asr}, T, last\_end\_idx)$  ▷ Fallback 2
13:    else
14:       $match \leftarrow match_{global}$  ▷ Use global match
15:    end if
16:  end if
17:  Append  $match$  to  $S_{aligned}$ 
18:   $last\_end\_idx \leftarrow match.end\_idx$  ▷ Update for next sequential search
19: end for
20: return  $S_{aligned}$ 

```

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

751 C Data Sources

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

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

758 All CSV files containing the source metadata are publicly available on Hugging Face at <https://huggingface.co/datasets/SamuelPfisterer1/EuroSpeech-Data-Sources>. These files
 759 provide complete transparency regarding the origins of our dataset and enable others to replicate or
 760 extend our data collection methodology.

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

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

D 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 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.