FalAR: A Large-scale Speaker-Annotated European Portuguese Speech Corpus of Parliamentary Sessions

Francisco Teixeira[1], **Carlos Carvalho**[1,2], **Mariana Julião**[1,2], **Catarina Botelho**[1], **Rubén Solera-Ureña**[1], **Sérgio Paulo**[1], **Thomas Rolland**[1], **Ben Peters**[1], **Isabel Trancoso**[1,2], **Alberto Abad**[1,2]
[1]INESC-ID, Lisbon, Portugal,
[2]Instituto Superior Técnico, Universidade de Lisboa, Portugal
{francisco.s.teixeira, carlos.carvalho}@inesc-id.pt

Abstract

State-of-the-art performance for Automatic Speech Recognition (ASR) largely depends on the availability of large-scale labeled corpora. This creates a demand for increased data collection efforts, particularly for under-represented languages and dialectal varieties. Due to having considerably fewer speakers (around 11 million), European Portuguese (EP) is overshadowed by Brazilian Portuguese (BP) (around 200 million speakers) in currently available large-scale speech data resources, resulting in under-performing speech-based systems for EP users. To address this gap, and following similar data collection efforts for other languages, we present FalAR, a large-scale, speaker-annotated speech corpus of European Portuguese parliamentary sessions. Spanning approximately 20 years, FalAR comprises 5,800 hours of speech data. In addition, 4,850 hours have speaker identity annotations, for a total of 1,180 speakers with associated metadata including age, gender, political affiliation, and parliamentary role. The corpus was built using a state-of-the-art EP CAMÕES ASR model for transcription-reference alignment. In this paper, we describe the data collection process, together with the main characteristics of the FalAR corpus. Furthermore, we evaluate the trade-off between data quantity and alignment accuracy on ASR performance, with our experiments demonstrating that incorporating FalAR as pre-training data yields up to 14% relative WER improvement over baseline models.

Keywords: parliamentary speech corpus, automatic speech recognition, speaker recognition

1. Introduction

Recent advances in Automatic Speech Recognition (ASR) have been driven by a combination of architectural innovations (Dong et al., 2018; Karita et al., 2019; Gulati et al., 2020; Kim et al., 2023; Rekesh et al., 2023), increased computational power, and the growing availability of large-scale labeled speech corpora (Chan et al., 2021; Radford et al., 2023; Kang et al., 2024; Peng et al., 2025). However, this trend inherently benefits majority languages or those with abundant online resources, while disadvantaging those with limited data (i.e., low-resource languages) (Chen et al., 2024). Furthermore, even among seemingly high-resource languages, certain dialects or regional varieties may remain under-represented. Varieties can differ substantially in pronunciation and vocabulary, causing models trained on the dominant variety to underperform when applied to others.

An example of this asymmetry can be found between European Portuguese (EP) and Brazilian Portuguese (BP), which are rarely differentiated in currently available speech databases, and are often treated as a completely homogeneous language. Therefore, speech recognition systems trained on large-scale web-scraped Portuguese data predominantly observe BP speech due to its substantially larger speaker base – approximately 197 million out of a total of 240 million Portuguese speakers – causing these models to perform suboptimally when applied to EP or other Portuguese varieties (Carvalho et al., 2025). A plausible explanation for this suboptimal performance lies in the phonetic, prosodic, syntactic and phonological differences between the two varieties. One of the most striking differences between BP and EP concerns vowel reduction, which is much more extreme in EP than in BP, but there are also notable differences at the syntactic and lexical levels (Mateus and d'Andrade, 2000; Rouas et al., 2008; Abad et al., 2009).

Even though EP has a much smaller speaker base than BP, it is nonetheless the native variety of around 11 million speakers. Ensuring fair and equitable access to speech technologies by EP speakers therefore demands the collection of large-scale resources for this variety. Historically, labeled resources for EP have been limited in scale, with the largest datasets falling short of 100 hours of speech data (Neto et al., 1997; Trancoso et al., 2003; Hagen and Neto, 2003). To date, one of the largest reported combinations of transcribed EP datasets corresponds to the training data of the CAMÕES models (Carvalho et al., 2025), amounting to only 425 hours of speech.

In this work, we leverage publicly available, manually annotated recordings of parliamentary sessions of the Portuguese parliament[1] to address

[1]https://www.parlamento.pt

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 5566–5577
11–16 May 2026.

this shortage. Specifically, we introduce FalAR,[2,3] a large-scale, speaker-annotated speech corpus containing 5,800 hours of transcription-reference aligned speech that spans approximately 20 years of parliamentary sessions of the Portuguese parliament. Alongside speech recordings and transcriptions, FalAR also includes detailed speaker identity and associated metadata, including age, gender, political affiliation and parliamentary role for 1,180 unique speakers, corresponding to more than 4,850 hours of speech.

We consider our work to be complimentary to the recently published EuroSpeech corpus (Pfisterer et al., 2025), which compiles speech data from 22 European parliaments, including the Portuguese parliament. The scale of EuroSpeech required its authors to apply a generic approach when collecting and annotating the data from each parliament, whereas our collection focused solely on the Portuguese parliament. This allowed us to annotate speaker identities and include speaker metadata, which EuroSpeech is lacking (resulting in speaker-dependent partitions), and to use the state-of-the-art EP ASR models of CAMÕES to generate pseudo-transcriptions for alignment, resulting in a larger and higher quality corpus.

In addition, we conduct a series of experiments that evaluate the performance of FalAR subsets with progressively higher alignment error rates, to analyze the trade-offs between data size and alignment accuracy, and their impact on training downstream ASR models for EP. When evaluated on the out-of-domain CAMÕES benchmark, performance of out-of-domain models trained with FalAR improves steadily as the amount of data – and correspondingly, the alignment error rate – increases, with results approaching the performance of baseline models trained with in-domain data. Moreover, employing FalAR as pre-training data prior to fine-tuning with in-domain speech allows for relative improvements of up to 14% WER compared to models trained from scratch.

Overall, the main contributions of this work are the following:

- We introduce FalAR, a large-scale, speaker-annotated speech corpus with 5,800 hours of aligned speech to address the shortage of speech resources for European Portuguese.
- We collect speaker identity and metadata, including age, gender, political affiliation and parliamentary role, for 4,850 hours of speech corresponding to 1,180 unique speakers, with longitudinal data spanning up to 20 years.
- We conduct several experiments to assess the trade-offs between data size and alignment accuracy, with our best models presenting a 14% WER relative improvement over our domain-specific baseline.

The paper is organized as follows: Section 2 presents the relevant related work; Section 3 describes the data collection effort and the FalAR corpus; Sections 4 and 5 detail this paper's experiments and results; finally, Section 6 presents conclusions and plans for future work.

2. Related Work

The development of speech resources for EP has always been closely linked to the development of EP ASR systems, as evidenced by early examples of data collection efforts for EP speech technologies. For instance, BD-PUBLICO (Neto et al., 1997), a corpus of 25 hours of read newspaper articles was collected to be used as the training data of an early large vocabulary hybrid Hidden Markov Model (HMM)/Deep Neural Network (DNN) system (Neto et al., 1998). Similarly, the EP portion of SpeechDat (Hagen and Neto, 2003), comprising 81 hours of narrow-band telephone speech, was collected as a part of an European-level collection of spoken language resources (Hoge et al., 1997) for the development of speech-based technologies. Likewise, ALERT (Trancoso et al., 2003), a broadcast news corpus comprising 74 hours of speech, was collected for and used to train the hybrid HMM/DNN AUDIMUS system, developed to automatically transcribe broadcast news in EP (Meinedo et al., 2001; Neto et al., 2008).

As ASR architectures became more data-demanding, research and development efforts in Portuguese progressively shifted towards BP, favoured by the availability of more extensive datasets resulting from its considerably larger speaker population (Alencar and Alcaim, 2008; Candido Junior et al., 2023; Lima et al., 2025; Evaldo Leal et al., 2025).

Contrarily, efforts for EP ASR became increasingly reliant on combinations of multiple datasets and small manually identified EP subsets of larger Portuguese corpora. For instance Carvalho (2021); Carvalho and Abad (2021) combined data from BD-PÚBLICO, ALERT and SpeechDat, amounting to close to 150 hours to train the first end-to-end Connectionist-Temporal-Classification (CTC)-Attention ASR model for EP. Campinho (2021) curated a similar amount of EP data (150 hours), leveraging 2 hours of manually identified EP speech from Common Voice (Ardila et al., 2020), 70

[2]The full corpus is available at https://huggingface.co/datasets/inesc-id/FalAR.

[3]In Portuguese, "falar" means "to speak"; AR is the acronym of "Assembleia da República", the official name of the Portuguese parliament.

hours from an in-house corpus sourced from the Portuguese RTP channel and 80 hours of Fala Bracarense (Centro de Estudos Humanísticos, Universidade do Minho, 2009), a corpus of sociolinguistic interviews, built for the study of the regional accent of the Braga region in Portugal. Mourão de Sá (2021) collected a total of 54 hours of EP speech, using around 33 hours from Europarl-ST (Iranzo-Sánchez et al., 2020) (a speech translation corpus of recordings from the European Parliament), and 21 hours from Multilingual-TEDx (a corpus of TEDx talks).

However, the models resulting from the aforementioned collection efforts were consistently outperformed by baseline hybrid HMM/DNN systems trained using the same data, evidencing the impact that the lack of resources has had in the development of end-to-end ASR systems for EP.

More recently, Carvalho et al. (2025) introduced CAMÕES,[4] an ASR framework for EP. It consists of an evaluation benchmark with 46 hours of EP data and a collection of state-of-the-art ASR models trained/fine-tuned with 425 hours of EP speech, compiled from a mix of proprietary corpora and publicly available sources, spanning multiple domains and demographic groups. This work was the first successful attempt to achieve EP results on par with the state-of-the-art for BP with end-to-end systems.

Portuguese speech subsets are also present in several well-known large-scale multilingual corpora. However, in most cases, the BP and EP are not differentiated. More importantly, BP is significantly more prevalent in these corpora. Corpora comprising EP speech include the above-mentioned CommonVoice (Ardila et al., 2020), with close to 2 hours of EP, Multilingual LibriSpeech (MLS) (Pratap et al., 2020), with around 56h of EP speech, and MuAViC (Anwar et al., 2023), reaching close to 20 hours of EP.[5] The YODAS dataset (Li et al., 2023; Peng et al., 2025) and the MOSEL collection (Gaido et al., 2024) include over 20,000h of Portuguese speech, sourced predominantly from YouTube, with a large proportion unlabeled or automatically labeled. However, the actual proportion of EP speech is unknown.

An increasingly common approach for creating large-scale speech and text resources has been the collection of recordings of parliamentary debates, which are often publicly available and accompanied by high-quality – although not always fully-verbatim – transcripts.

Examples of text-only resources include ParlaMint (Erjavec et al., 2023), a corpus of transcripts of parliamentary proceedings of 26 national European parliaments, and Europarl (Koehn, 2005), a translation corpus that includes parallel text from the European Parliament in 11 languages. Textual corpora from the Portuguese parliament have already been released in PTPARL-D (Almeida et al., 2021) and ParlaMint-PT (Aires et al., 2024), with transcriptions of debates spanning 1976 to 2019 and 2005 to 2019, respectively.

In addition, there is a growing number of speech corpora compiled from parliamentary data. These include large-scale corpora, such as Europarl-ASR (Garcés Díaz-Munío et al., 2021) for English as well as corpora for under-represented languages such as Danish (Kirkedal et al., 2020), Catalan (Kulebi et al., 2022), Norwegian (Solberg and Ortiz, 2022), Croatian, Polish, and Serbian (Ljubešić et al., 2024).

To date, the multilingual VoxPopuli corpus, which contains recordings from the European Parliament representing all languages in the European Union, is the largest available resource for EP, with 17.5k hours of unlabeled EP speech that can be used for self-supervised learning (SSL)-based pretraining (Mohamed et al., 2022; Zhang et al., 2023). Nevertheless, achieving robust performance in downstream tasks still requires labeled data that span a wide range of speech domains, age groups, and other relevant speech factors.

More recently, the EuroSpeech corpus (Pfisterer et al., 2025) has released over 78k hours of labeled speech from 22 European national parliaments, including around 5,100 hours from the Portuguese parliament. Our work complements this work by focusing specifically on the collection of European Portuguese parliamentary speech. The use of the CAMÕES state-of-the-art ASR model allowed us to generate pseudo-transcripts for the alignment of speech and the parliamentary proceedings transcripts, resulting in higher quality annotations. Furthermore, we include manually verified speaker identity annotations and speaker metadata, information that is lacking from the EuroSpeech, making its train/dev/test partitions speaker-dependent.

3. FalAR

The main objective of this work is to build a large-scale speech corpus for EP, leveraging the publicly available video recordings of Portuguese parliamentary meetings and corresponding manual transcriptions.

To achieve this, we collected the recordings, extracted and segmented the audio signals, and generated automatic transcriptions. These transcriptions were then aligned with the reference texts to determine the text segments corresponding to each utterance, which served as the ground-truth labels.

[4] https://huggingface.co/datasets/inesc-id/camoes_asr

[5] EP hours were determined by manual inspection and annotation.

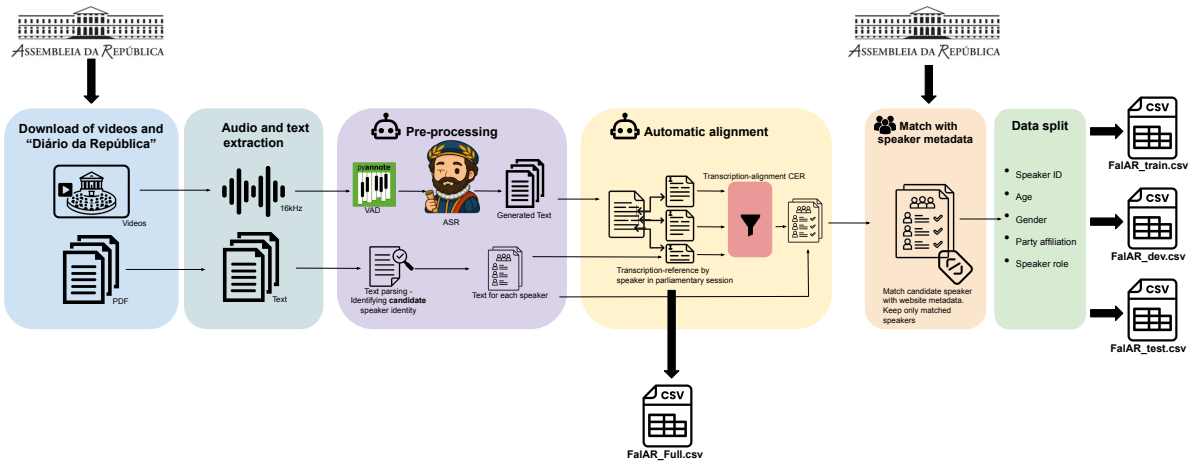


Figure 1: FalAR data collection and processing pipeline.

For each utterance, we further identified a candidate speaker and matched them to the recording's metadata to obtain reliable speaker labels. Finally, for every identified speaker, we manually annotated gender and date of birth (when this information was available online). The full data collection and processing pipeline is represented in Figure 1.

The remainder of this section provides detailed description of the data collection, alignment, and annotation processes, as well as an overview of the resulting corpus.

3.1. Data collection

Since 2005, the Portuguese parliament has made its plenary sessions publicly accessible through its Audio-Visual Archive (AVA),[6] which contains complete session recordings as well as shorter clips of individual interventions. This archive is complemented by "Diário da Assembleia da República" (DAR)[7] – the official gazette of the Portuguese parliament – which provides detailed session summaries and full transcriptions (although not necessarily verbatim) of all interventions from each parliamentary meeting.

To construct our corpus, we first identified all available legislatures, legislative sessions, parliamentary meetings and individual interventions, and compiled a list of intervention video URLs. In parallel, we compiled a list of URLs for the DARs corresponding to each parliamentary meeting. This semi-automatic process resulted in the identification and successful download of 104,031 video interventions, along with their associated metadata (including speaker name, political affiliation, topic of intervention and parliamentary role) in HTML format, covering 2,041 parliamentary sessions over the last 20 years. For each of these sessions, we also retrieved the corresponding DARs, which include session summaries and full transcriptions of all interventions.

3.2. Data pre-processing

3.2.1. Audio

Audio was extracted from all downloaded videos, converted to 16-bit PCM format, and downsampled to 16 kHz. Each audio file was segmented into utterances of up to 30 seconds, using Pyannote's Voice Activity Detection (VAD) system (Bredin, 2023). Subsequently, each segment was transcribed using CAMÕES's WhisperLv3-X[8].

3.2.2. Text

The downloaded DARs in HTML format were converted into plain text files, and each file underwent a semi-automatic filtering process to produce text files containing only the relevant content of speaker interventions (i.e., any spoken utterance by a participant during a parliamentary session that has been transcribed).

Following this initial pre-processing stage, the next step involved identifying the beginning of each intervention, indicated by the pattern: "{speaker name}({affiliation}): - ". Detecting this pattern enabled the extraction of the text between successive occurrences and its subsequent assignment to a candidate speaker, a necessary step to ensure correct speaker attribution.

However, manual inspection of initial groupings identified several cases in which speakers were

[6] https://av.parlamento.pt

[7] https://debates.parlamento.pt/

[8] https://huggingface.co/inesc-id/WhisperLv3-EP-X a version of Whisper (Radford et al., 2023) fine-tuned on 425 hours of EP speech (Carvalho et al., 2025).

designated by "O(A) Orador(a)" (i.e., *the speaker*), a short-hand notation in DAR to indicate the main speaker of an intervention, when that speaker has been interrupted by another representative. To correctly identify as many speakers as possible, we added a heuristic rule whereby, if the speaker is identified as "O(A) Orador(a)", the next-to-last speaker not labeled as "O(A) Orador(a)" is selected as the candidate speaker.

3.3. Transcription-reference alignment

After pre-processing the audio and text data, we aligned the automatic transcriptions with the reference DAR texts. To accomplish this, we used the Smith-Waterman algorithm (Smith and Waterman, 1981), which performs local sequence alignment by identifying the contiguous region within each reference text that most closely corresponds to a given transcription segment, accounting for partial matches and gaps.

To preserve speaker candidate information, each segment was aligned with portions of the reference text grouped by candidate speaker. For each segment, the aligned reference segment (and corresponding speaker) with the lowest Character Error Rate (CER) relative to the automatic segment transcription was selected as the gold-standard label. This is done for two reasons: first, we consider that the segment from the reference text has a higher likelihood of being the true text; and second, selecting the reference text as the gold-standard label allows us to keep punctuation and capitalization, something that the automatically generated transcriptions do not contain.

3.4. Speaker annotation

The process described in the previous section resulted in the identification of 3,055 candidate speakers. Each speaker name was manually inspected to remove duplicate names due to typos, annotation errors and formatting issues. In addition, a number of speakers were identified only by their office (e.g., "minister", or "secretary of state"), in which case, they were annotated with the true speaker's name. This process resulted in 1,200 identified candidate speakers.

To minimize potential errors caused by alignment or formatting issues, we only considered as correct, candidate speakers whose names matched the speaker name contained in the videos' metadata. It is important to note that this decision is particularly strict, since a large majority of videos contain interventions from more than one speaker, often including not only the speaker annotated in the video's metadata, but also a short intervention by the president of the parliament giving the floor to the speaker. Consequently, a large number of such interventions were left as non-verified speakers, i.e., stored without speaker annotation and left out of the speaker-independent partitions.

Following this stage, we proceeded to annotate each speaker with gender and date of birth, using online resources. Out of the total 1,200 speakers, we were able to annotate the dates of birth (or, in a small number of cases, the year of birth) of 1,180 speakers. All identified speakers were annotated with gender information.

To protect the privacy of the speakers present in this corpus to the best of our ability, we only provide numeric speaker identifiers, and omit the speaker's names. Moreover, we provide age annotations at the utterance level, instead of the speakers' dates of birth.

3.5. Corpus description

The processes outlined in the previous sections yielded a corpus comprising 5,799 hours of transcribed speech data, of which 4,852 hours are annotated with speaker age and gender information, as summarized in Table 1.

For each audio utterance, we provide (1) the reference segment transcript obtained with the Smith-Waterman alignment algorithm, (2) the automatic transcription generated using CAMÕES's WhisperLv3-X, (3) the CER between the two transcriptions, (4) the date of recording, (5) speaker id, (6) speaker gender, (7) speaker age, (8) speaker political affiliation, (9) speaker role – e.g. *Presidente da República* (President of the Republic), *Deputado* (Member of parliament) –, and (10) intervention topic.

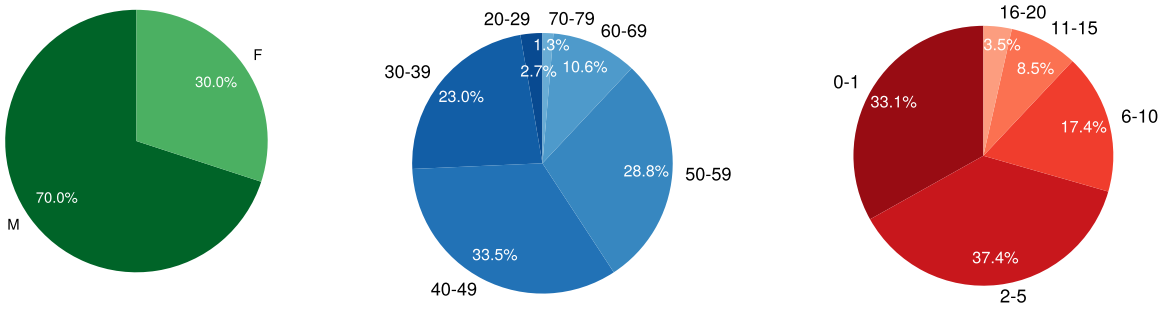
Table 1 also reports the size of the corpus across varying CER thresholds, obtained for the transcription-reference alignments. Lower CER thresholds indicate higher confidence in the accuracy of the aligned DAR transcription.

The subset of the corpus containing speaker-level information includes 1,200 unique speakers, all of which have been annotated with gender information, while only 1,180 have been annotated with age information. Approximately 70% of the speakers are male, while the remaining 30% are female. The speakers' ages range from 20 to 79 years, with a predominantly middle-aged population. Each speaker contributes an average of four hours of speech. About 70% of the speakers appear in recordings spanning up to five years, while approximately 3.5% of the speakers have recordings that extend over a period of 16 to 20 years. More detailed information regarding the demographic distribution of the corpus is presented in Figure 2.

Besides making the entire dataset available (*FalAR_Full.csv* in Fig. 1), we also provide standardized speaker-independent train, development

	All data		Data with speaker age and gender information						
	Duration	Word tokens	Duration (hours)			Speaker count			Word Tokens
CER	(hours)	(millions)	Total	M	F	Total	M	F	(millions)
<5%	1,503	13	1,366	923	443	1,159	772	387	12
<10%	2,598	23	2,370	1,621	749	1,173	783	390	21
<15%	3,422	31	3,122	2,153	969	1,177	787	390	28
<20%	4,026	36	3,664	2,539	1,125	1,178	788	390	33
Total	5,799	52	4,852	3,399	1,453	1,180	790	390	44

Table 1: Dataset description, as a function of the CER threshold. *M* and *F* refer to male and female.



(a) Gender (b) Age (c) Longitudinal intervals

Figure 2: FalAR demographic distribution.

and test splits for the subset of the corpus containing speaker-level information, to promote reproducible research with this corpus. These partitions correspond to *FalAR_train.csv*, *FalAR_dev.csv*, and *FalAR_test.csv* in Figure 1 and comprise 4662/1171, 40/12, and 34/17 (hours/speakers), respectively. The speakers in each partition are randomly and uniformly selected from the full set of annotated speakers.

Note that the test set is exclusively comprised of utterances with a maximum alignment CER of 5%, to privilege high-quality transcriptions. This decision may make the test set less challenging, since utterances with lower alignment error rates may be inherently easier to transcribe, as they correspond to sentences where the model used to generate the transcriptions matched the reference text more closely. Nevertheless, we consider ensuring high-quality transcriptions in the test set to be more important than creating a more challenging but less reliable test set.

4. Experimental Setup

4.1. Data

To assess the impact of the proposed corpus on the performance of ASR models for European Portuguese, we conduct a series of experiments using different data configurations.

First, to determine the impact that different alignment error rates have in downstream ASR systems, we prepare five subsets of FalAR to train corresponding ASR models, as presented in Table 1, namely FalAR <{5,10,15,20}% CER and FalAR <20% CER + WL (all denoted by $FalAR_{<x\%}$ from hereon). The latter corresponds to the full corpus, where the labels for speech segments with CER $\geq$ 20% correspond to the automatically generated transcripts (WL - weakly supervised), using the assumption that these are more reliable than the segments obtained by the alignment algorithm. The data used in these five partitions is taken from *All data* (cf. Table 1), for which speaker information may not necessarily be available, to maximize the quantity of data available. The models trained with the aforementioned sets are all evaluated with in-domain data, using FalAR's test set (34.5h of speech with CER $<$ 5%).

To explore the out-of-domain performance of the above-mentioned models, and to understand the impact of using FalAR as a pre-training corpus, followed by fine-tuning with data matching the evaluation domain, we used the corpus collected by the authors of CAMÕES (Carvalho et al., 2025). This dataset comprises over 470h of speech compiled from a set of 18 corpora from multiple domains – 14 of these datasets are used for training and another intersecting set of 14 datasets is used in the test benchmark. The training set (denoted as EP-

425) is 425 hours-long, whereas the test set is 46 hours-long, both comprising five domains, namely, read speech (RS), broadcast news (BN), talks and lectures (T/L), conversational speech (CS), and sociolinguistic interviews (SI). An ASR model was additionally trained using solely CAMÕES, to provide a baseline with which to compare the FalAR-based models to.

It is important to note that, since we are using the *All data* set for training, there may exist some speaker overlap with the test partition for non-speaker annotated samples. This is particularly true for $FalAR_{<20\%+WL}$, where the difference between the speaker annotated and non-speaker annotated sets is close to 1,000 hours of speech. This decision balanced the goal of understanding the impact of larger-scale data on out-of-domain performance against the computational cost of rerunning the full set of experiments with our own speaker-independent partitions to better analyse the same effects on in-domain parliamentary data.

4.2. Implementation

We used the ESPnet toolkit (Watanabe et al., 2018) for the core implementation and evaluation of our work. More specifically, we followed the LibriSpeech 960 (Panayotov et al., 2015) recipe in ESPnet for training, decoding, and evaluation. All evaluated ASR models correspond to an E-Branchformer (Kim et al., 2023) with 144M trainable parameters, using 8x downsampling (Rekesh et al., 2023) and Flash Attention (Dao et al., 2022) to improve training and inference efficiency. The model's encoder comprises 17 layers, whereas the decoder is a 6-layer Transformer, both adapted from the original recipe. For the encoder module, we applied Rotary Positional Embeddings (RoPE) (Su et al., 2024). We also adopted a piecewise-linear learning rate schedule (Peng et al., 2024), gradually increasing the learning rate from 2.0e-4 over the initial 15k steps to 2.0e-3 over the next 20k steps.

The training of the models with the different FalAR CER subsets in Table 1 was carried out on a single NVIDIA H200 NVL GPU with 144GB of memory, using 120M batch bins and 15 epochs, except for $FalAR_{<20\%}$ and $FalAR_{<20\%+WL}$ models, which were trained for only 10 epochs. Training and fine-tuning of models with the CAMÕES corpus was performed on a single NVIDIA RTX 6000 Ada generation GPU with 48GB of memory, using 14M batch bins and 35 epochs.

All models use the same text normalizer, based on the standard procedures used by Whisper, removing, among others, punctuation and capitalization, and expanding common acronyms. Performances are reported using WER (%). Utterances shorter than 0.1 seconds or longer than 30 seconds were excluded during training and fine-tuning.

5. Results

5.1. In-domain performance

Table 2 presents the in-domain results for the FalAR test set together with the out-of-domain performance on the CAMÕES benchmark, evaluated across its five domains.

For the FalAR test set, we observe that performance generally improves as the training data size increases, with the largest gain and best overall performance of 3.1% WER being caused by the addition of the WL data ($FalAR_{<20\%+WL}$ set). However, the second best result of 4.2% WER comes from the $FalAR_{<5\%}$ set, worsening to 6.7% when the $FalAR_{<10\%}$ set is used, and only then steadily dropping. These results for in-domain data suggest advantages of using either smaller training sets with high-quality alignments ($FalAR_{<5\%}$) or very large training sets, even with a lower quality ($FalAR_{<20\%+WL}$), compared to medium-sized training sets with intermediate transcription-reference alignment error levels. However, it is also possible that there exists a larger speaker overlap between the training and test data for the $FalAR_{<20\%+WL}$ subset that explains the strong performance of this model.

5.2. Out-of-domain performance

From the high-quality but smaller $FalAR_{<5\%}$ training set, to the larger $FalAR_{<20\%+WL}$ set, the average out-of-domain performance on CAMÕES improves from 31.1% to 24.4% WER. This proves that increasing the size of the FalAR training dataset allows the model to generalise better, even with a larger prevalence of transcription errors. Particularly, we observe that the addition of the WL data ($\sim$1800 additional hours) strongly contributes to the improvement of the out-of-domain performance of the model. When trained with this data, the model starts to become competitive with the baseline model trained with the CAMÕES EP-425 set, achieving comparable or even stronger results in the T/L and CS domains. We hypothesize that sentences with higher alignment errors are harder for ASR models to transcribe than those with low alignment error rates due to worse and/or more varied acoustic conditions. Therefore, even though their transcriptions contain more errors, including these utterances in training helps the model generalize better to other domains.

5.3. Fine-tuned performance

The results obtained after fine-tuning the pre-trained models using 425 hours of speech data from the CAMÕES framework (EP-425) are presented in the lower half of Table 2.

Pre-training data	Fine-tuning data	FalAR	CAMÕES					
			RS	BN	T/L	CS	SI	Avg.
EP-425	–	15.5	12.8	8.00	21.2	22.1	39.1	20.6
$FalAR_{<5\%}$	–	4.2	25.8	13.6	27.5	28.5	60.2	31.1
$FalAR_{<10\%}$	–	6.7	25.5	13.5	28.6	30.1	61.3	31.8
$FalAR_{<15\%}$	–	5.0	24.6	12.2	25.5	26.3	57.0	29.1
$FalAR_{<20\%}$	–	5.1	24.1	12.4	26.3	27.2	58.4	29.7
$FalAR_{<20\%+WL}$	–	3.1	19.0	9.9	20.9	22.2	50.0	24.4
$FalAR_{<5\%}$	EP-425	13.9	11.1	7.8	20.5	20.0	37.3	19.3
$FalAR_{<10\%}$	EP-425	13.2	10.5	6.8	19.3	19.5	36.0	18.4
$FalAR_{<15\%}$	EP-425	12.8	9.4	6.6	18.2	18.8	35.7	17.7
$FalAR_{<20\%}$	EP-425	13.5	9.6	6.7	19.6	19.9	37.2	18.6
$FalAR_{<20\%+WL}$	EP-425	11.7	9.9	6.7	18.2	18.5	35.2	17.7

Table 2: WER (%) on FalAR test set and CAMÕES benchmark.

As anticipated, fine-tuning markedly improves performance on the CAMÕES benchmark relative to the out-of-domain condition. Furthermore, all models incorporating pre-training show superior performance compared to the baseline model trained solely on EP-425, even though the pre-training data domain is highly specific (parliamentary speech) and differs substantially from those represented in the CAMÕES benchmark.

Up to the 15% CER threshold ($FalAR_{<15\%}$ set), larger amounts of pre-training data lead to better performance, with the best average result presenting a relative improvement of 14% w.r.t. to the baseline model. However, beyond this point, model performance first degrades, and then stabilizes with the addition of the WL data. This is opposed to what was observed when evaluating the same models in the out-of-domain setting. This may indicate that there is a maximum amount of pre-training data from the same domain from which models are able to benefit when fine-tuned for downstream domains. Alternately, as the model size remained fixed as we increased the amount of training data, the model pre-trained with $FalAR_{<20\%+WL}$ may simply not have the capacity to retain enough information about its additional pre-training data for it to still be beneficial after fine-tuning. The fact that the fine-tuned models' performance on FalAR's test set is markedly worse than the pre-trained models' performance on this data may provide evidence to this argument. Therefore, these results demand further experimentation to ascertain the exact causes of the reported performance.

6. Conclusions

This work introduces FalAR, to the best of our knowledge, the largest publicly available annotated European Portuguese speech corpus, totalling 5,800 hours of parliamentary speech data. Our results show that using FalAR as pre-training data followed by in-domain fine-tuning improves ASR performance across all domains of the CAMÕES benchmark when compared to a strictly in-domain baseline model.

Although we provide standardized speaker-independent partitions together with this corpus, we do not provide results for the speaker-independent setup. We aim to address this in future work.

Furthermore, the proposed large-scale corpus offers two key advantages. First, it is expandable, as new recordings and corresponding transcripts from Portuguese parliamentary sessions are released continuously. Second, this corpus contains rich speaker metadata, including speaker identity, age, gender, and party affiliation, for 4,850 hours of speech from 1,180 speakers. These annotations do not only allow fair and controlled data splits for model development, but also longitudinal studies of speech and speaker traits spanning two decades. This dataset may thus serve as a valuable resource for research on ageing, charisma and other behavioural attributes, political science and possibly a wide range of unforeseen applications.

Future improvements to the FalAR corpus may also focus on providing baseline results for speaker recognition, punctuated and capitalized ASR models, improving high alignment CER transcriptions through weakly supervised methods, improving speaker labels through automatic speaker recognition, and releasing video data to support broader multimodal research. In a multimodal format, this resource could also potentially be turned into a searchable archive of parliamentary sessions (Hansen et al., 2005).

7. Ethical considerations and limitations

The source data that we curated and analysed to compile FalAR was obtained from publicly available open data resources (see Section 3.1).

In releasing the accompanying metadata, we deliberately omit personally identifiable information such as speaker names and dates of birth, and instead provide anonymised speaker identifiers and age information. Nevertheless, it must be acknowledged that, even if discarding biometric re-identification, complete anonymisation cannot be guaranteed, as re-identification is still possible by cross-referencing the provided metadata with publicly accessible information in the Portuguese Parliament website. A further limitation of the data curation process rises from the extraction of speaker identifiers from PDF documents, which were frequently inconsistently formatted. These inconsistencies occasionally led to incomplete metadata entries.

Finally, another limitation concerns the absence of punctuation marks in the automatically generated transcriptions, which may affect the quality of the alignment between the transcriptions and the reference texts. Nevertheless, the aligned reference transcriptions retain punctuation, offering an advantage for future research based on this corpus.

8. Acknowledgements

Work supported by Portuguese national funds through Fundação para a Ciência e a Tecnologia, I.P. (FCT) under projects UID/50021/2025 (DOI: https://doi.org/10.54499/UID/50021/2025) and UID/PRR/50021/2025 (DOI: https://doi.org/10.54499/UID/PRR/50021/2025) and by the Portuguese Recovery and Resilience Plan and NextGenerationEU European Union funds under project C644865762-00000008 (ACCELERAT.AI).

9. Bibliographical References

Alberto Abad, Isabel Trancoso, Nelson Neto, and M. Céu Viana. 2009. Porting an european portuguese broadcast news recognition system to brazilian portuguese. In *Interspeech 2009*, pages 92–95.

José Aires, Aida Cardoso, Rui Pereira, and Amália Mendes. 2024. Compiling and exploring a Portuguese parliamentary corpus: ParlaMint-PT. In *Proc. of the IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora (ParlaCLARIN)@ LREC-COLING 2024*, pages 12–20.

Vladimir Fabregas Surigué de Alencar and Abraham Alcaim. 2008. LSF and LPC-derived features for large vocabulary distributed continuous speech recognition in Brazilian Portuguese. In *Proc. 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1237–1241.

Paulo Almeida, Manuel Marques-Pita, and Joana Gonçalves-Sá. 2021. PTPARL-D: an annotated corpus of forty-four years of Portuguese parliamentary debates. *Corpora*, 16(3):337–348.

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. In *Proc. Interspeech*, pages 4064–4068.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common Voice: A Massively-Multilingual Speech Corpus. In *Proc. LREC*, pages 4218–4222.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *Proc. Interspeech 2023*.

Adriano Campinho. 2021. Automatic speech recognition for European Portuguese. Master's thesis, Escola de Engenharia, Universidade do Minho, Braga, Portugal, July. Available at https://hdl.handle.net/1822/78249.

Arnaldo Candido Junior, Edresson Casanova, Anderson Soares, Frederico Santos de Oliveira, Lucas Oliveira, Ricardo Corso Fernandes Junior, Daniel Peixoto Pinto da Silva, Fernando Gorgulho Fayet, Bruno Baldissera Carlotto, Lucas Rafael Stefanel Gris, and Sandra M. Aluísio. 2023. CORAA ASR: a large corpus of spontaneous and prepared speech manually validated for speech recognition in Brazilian Portuguese. *Language Resources and Evaluation*, 57:1139–1171.

Carlos Carvalho. 2021. TRIBUS: An end-to-end automatic speech recognition system for European Portuguese. Master's thesis, Instituto Superior Técnico, Universidade de Lisboa, Lisbon, Portugal, January. Available at https://fenix.tecnico.ulisboa.pt/downloadFile/1126295043839127/81395-carlos-carvalho_dissertacao.pdf.

Carlos Carvalho and Alberto Abad. 2021. TRIBUS: An end-to-end automatic speech recognition system for European Portuguese. In *Proc. IberSPEECH*, pages 185–189.

Carlos Carvalho, Francisco Teixeira, Catarina Botelho, Anna Pompili, Rubén Solera-Ureña, Sérgio Paulo, Mariana Julião, Thomas Rolland, John Mendonça, Diogo Pereira, Isabel Trancoso, and Alberto Abad. 2025. CAMÕES: A comprehensive automatic speech recognition benchmark for European Portuguese. In *Accepted to ASRU*.

Centro de Estudos Humanísticos, Universidade do Minho. 2009. Perfil Sociolinguístico da Fala Bracarense. https://sites.google.com/site/projectofalabracarense/. Accessed: 2025-10-24.

William Chan, Daniel Park, Chris Lee, Yu Zhang, Quoc Le, and Mohammad Norouzi. 2021. Speechstew: Simply mix all available speech recognition data to train one large neural network. *arXiv preprint, https://arxiv.org/abs/2104.02133*.

William Chen, Wangyou Zhang, Yifan Peng, Xinjian Li, Jinchuan Tian, Jiatong Shi, Xuankai Chang, Soumi Maiti, Karen Livescu, and Shinji Watanabe. 2024. Towards robust speech representation learning for thousands of languages. *arXiv pre-print, https://arxiv.org/abs/2407.00837*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Proc. NEURIPS*, pages 16344 – 16359.

Linhao Dong, Shuang Xu, and Bo Xu. 2018. Speech-transformer: A no-recurrence sequence-to-sequence model for speech recognition. In *Proc. ICASSP*, pages 5884–5888.

Tomaž Erjavec, Maciej Ogrodniczuk, Petya Osenova, Nikola Ljubešić, Kiril Simov, Andrej Pančur, Michał Rudolf, Matyáš Kopp, Starkaður Barkarson, Steinþór Steingrímsson, Çağrı Çöltekin, Jesse de Does, Katrien Depuydt, Tommaso Agnoloni, Giulia Venturi, María Calzada Pérez, Luciana D. de Macedo, Costanza Navarretta, Giancarlo Luxardo, Matthew Coole, Paul Rayson, Vaidas Morkevičius, Tomas Krilavičius, Roberts Darǵis, Orsolya Ring, Ruben van Heusden, Maarten Marx, and Darja Fišer. 2023. The ParlaMint corpora of parliamentary proceedings. *Language Resources and Evaluation*, 57(1):415–448.

Sidney Evaldo Leal, Arnaldo Candido Junior, Ricardo Marcacini, Edresson Casanova, Odilon Gonçalves, Anderson Silva Soares, Rodrigo Freitas Lima, Lucas Rafael Stefanel Gris, and Sandra Aluísio. 2025. MuPe life stories dataset: Spontaneous speech in Brazilian Portuguese with a case study evaluation on ASR bias against speakers groups and topic modeling. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6076–6087, Abu Dhabi, UAE. Association for Computational Linguistics.

Marco Gaido, Sara Papi, Luisa Bentivogli, Alessio Brutti, Mauro Cettolo, Roberto Gretter, Marco Matassoni, Mohamed Nabih, and Matteo Negri. 2024. MOSEL: 950,000 hours of speech data for open-source speech foundation model training on EU languages. In *Proc. EMNLP*, pages 13934–13947, Miami, Florida, USA. Association for Computational Linguistics.

Gonçal Garcés Díaz-Munío, Joan Albert Silvestre Cerdà, Javier Jorge-Cano, Adrián Giménez Pastor, Javier Iranzo-Sánchez, Pau Baquero-Arnal, Nahuel Roselló, Alejandro Manuel Pérez-González de Martos, Jorge Civera Saiz, José Alberto Sanchis Navarro, and Juan Alfons. 2021. Europarl-ASR: A large corpus of parliamentary debates for streaming ASR benchmarking and speech data filtering/verbatimization. *Proc. Interspeech 2021*, pages 3695–3699.

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*.

Astrid Hagen and Joao P Neto. 2003. HMM/MLP hybrid speech recognizer for the Portuguese telephone SpeechDat corpus. In *Proc. PROPOR*, pages 126–134.

J.H.L. Hansen, Rongqing Huang, Bowen Zhou, M. Seadle, J.R. Deller, A.R. Gurijala, M. Kurimo, and P. Angkititrakul. 2005. Speechfind: Advances in spoken document retrieval for a national gallery of the spoken word. *IEEE Transactions on Speech and Audio Processing*, 13(5):712–730.

Harald Hoge, Herbert S Tropf, Richard Winski, Henk van den Heuvel, Reinhold Haeb-Umbach, and Khalid Choukri. 1997. European speech databases for telephone applications. In *Proc. ICASSP*, volume 3, pages 1771–1774. IEEE.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià

Giménez, Albert Sanchis, Jorge Civera, and Alfons Juan. 2020. Europarl-ST: A multilingual corpus for speech translation of parliamentary debates. In *Proc. ICASSP 2020*, pages 8229–8233.

Wei Kang, Xiaoyu Yang, Zengwei Yao, Fangjun Kuang, Yifan Yang, Liyong Guo, Long Lin, and Daniel Povey. 2024. Libriheavy: a 50,000 hours asr corpus with punctuation casing and context. In *Proc. ICASSP*.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, Shinji Watanabe, Takenori Yoshimura, and Wangyou Zhang. 2019. A comparative study on transformer vs RNN in speech applications. In *Proc. ASRU*.

Kwangyoun Kim, Felix Wu, Yifan Peng, Jing Pan, Prashant Sridhar, Kyu J. Han, and Shinji Watanabe. 2023. E-Branchformer: Branchformer with enhanced merging for speech recognition. In *Proc. SLT*.

Andreas Kirkedal, Marija Stepanović, and Barbara Plank. 2020. FT Speech: Danish parliament speech corpus. In *Proc. Interspeech*, pages 442–446.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Baybars Kulebi, Carme Armentano-Oller, Carlos Rodriguez-Penagos, and Marta Villegas. 2022. ParlamentParla: A speech corpus of Catalan parliamentary sessions. In *Proceedings of the Workshop ParlaCLARIN III within the 13th Language Resources and Evaluation Conference*, pages 125–130, Marseille, France. European Language Resources Association.

Xinjian Li, Shinnosuke Takamichi, Takaaki Saeki, William Chen, Sayaka Shiota, and Shinji Watanabe. 2023. Yodas: Youtube-oriented dataset for audio and speech. In *Proc. ASRU*, pages 1–8.

Rodrigo Lima, Sidney E. Leal, Arnaldo Candido Junior, and Sandra M. Aluísio. 2025. A large dataset of spontaneous speech with the accent spoken in São Paulo for automatic speech recognition evaluation. In *Proc. Intelligent Systems: 34th Brazilian Conference (BRACIS)*, pages 33–47.

Nikola Ljubešić, Peter Rupnik, and Danijel Koržinek. 2024. The parlaspeech collection of automatically generated speech and text datasets from parliamentary proceedings. In *International Conference on Speech and Computer*, pages 137–150. Springer.

Maria Helena Mateus and Ernesto d'Andrade. 2000. *The Phonology Of Portuguese*. Oxford University Press.

Hugo Meinedo, Nuno. Souto, and João P. Neto. 2001. Speech recognition of broadcast news for the European Portuguese language. In *Proc. ASRU*, pages 319–322.

Abdelrahman Mohamed, Hung-yi Lee, Lasse Borgholt, Jakob D. Havtorn, Joakim Edin, Christian Igel, Katrin Kirchhoff, Shang-Wen Li, Karen Livescu, Lars Maaløe, Tara N. Sainath, and Shinji Watanabe. 2022. Self-supervised speech representation learning: A review. *IEEE Journal of Selected Topics in Signal Processing*, 16(6):1179–1210.

João Manuel Alves Mourão de Sá. 2021. Reconhecimento de fala em português de Portugal num contexto com poucos recursos. Master's thesis, Faculdade de Ciências, Universidade do Porto, Porto, Portugal, November. Available at `https://hdl.handle.net/10216/139258`.

Joao P. Neto, Ciro Martins, and Luis B. Almeida. 1998. A large vocabulary continuous speech recognition hybrid system for the portuguese language. In *5th International Conference on Spoken Language Processing (ICSLP 1998)*, page paper 0562.

João P. Neto, Ciro Martins, Hugo Meinedo, and Luis B Almeida. 1997. The design of a large vocabulary speech corpus for Portuguese. In *Proc. Eurospeech*, pages 1707–1710.

João P. Neto, Hugo Meinedo, Márcio Viveiros, Renato Cassaca, Ciro Martins, and Diamantino Caseiro. 2008. Broadcast news subtitling system in Portuguese. In *Proc. ICASSP*, pages 1561–1564.

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

Yifan Peng, Shakeel Muhammad, Yui Sudo, William Chen, Jinchuan Tian, Chyi-Jiunn Lin, and Shinji Watanabe. 2025. OWSM v4: Improving open whisper-style speech models via data scaling and cleaning. In *Proc Interspeech*.

Yifan Peng, Yui Sudo, Muhammad Shakeel, and Shinji Watanabe. 2024. OWSM-CTC: An Open Encoder-Only Speech Foundation Model for Speech Recognition, Translation, and Language

Identification. In *Proc. ACL (Volume 1: Long Papers)*, pages 10192–10209.

Samuel Pfisterer, Florian Grötschla, Luca Lanzendörfer, Florian Yan, and Roger Wattenhofer. 2025. Eurospeech: A multilingual speech corpus. *Proc. NeurIPS*.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. 2020. Mls: A large-scale multilingual dataset for speech research. In *Proc. Interspeech*, 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 *Proc. ICML*.

Dima Rekesh, Nithin Rao Koluguri, Samuel Kriman, Somshubra Majumdar, Vahid Noroozi, He Huang, Oleksii Hrinchuk, Krishna Puvvada, Ankur Kumar, Jagadeesh Balam, and Boris Ginsburg. 2023. Fast conformer with linearly scalable attention for efficient speech recognition. In *Proc. ASRU*, pages 1–8.

Jean-Luc Rouas, Isabel Trancoso, Céu Viana, and Mónica Abreu. 2008. Language and variety verification on broadcast news for portuguese. *Speech Communication*, 50(11):965–979.

Temple F. Smith and Michael S. Waterman. 1981. Identification of common molecular subsequences. *Journal of Molecular Biology*, 147(1):195–197.

Per Erik Solberg and Pablo Ortiz. 2022. The Norwegian parliamentary speech corpus. In *Proc. LREC*, pages 1003–1008, Marseille, France. European Language Resources Association.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. RoFormer: Enhanced transformer with Rotary Position Embedding. *Neurocomputing*, 568:127063.

Isabel Trancoso, Joao P. Neto, Hugo Meinedo, and Rui Amaral. 2003. Evaluation of an alert system for selective dissemination of broadcast news. In *Proc. Interspeech*, pages 1257–1260.

Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai. 2018. Espnet: End-to-end speech processing toolkit. In *Proc. Interspeech*.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, Zhong Meng, Ke Hu, Andrew Rosenberg, Rohit Prabhavalkar, Daniel S. Park, Parisa Haghani, Jason Riesa, Ginger Perng, Hagen Soltau, Trevor Strohman, Bhuvana Ramabhadran, Tara Sainath, Pedro Moreno, Chung-Cheng Chiu, Johan Schalkwyk, Françoise Beaufays, and Yonghui Wu. 2023. Google USM: Scaling Automatic Speech Recognition Beyond 100 Languages. *arXiv pre-print, https://arxiv.org/abs/2303.01037*.