FPSC: A Sustainable Pipeline for Building a Faroese Parliamentary Speech Corpus

Dávid í Lág[1], **Barbara Scalvini**[1], **Carlos Mena**[2], **Jón Guðnason**[3]
[1]Faculty of Science and Technology, University of the Faroe Islands, Faroe Islands
[2]Language Technologies Laboratory, Barcelona Supercomputing Center (BSC), Spain
[3]Department of Engineering, Reykjavik University, Iceland
{davidl, barbaras}@setur.fo, carlos.hernandez@bsc.es, jg@ru.is

Abstract

This work addresses the lack of large-scale, natural speech data for Faroese automatic speech recognition. Existing resources, such as the 100-hour *Ravnursson* corpus, consist of read speech and do not capture the spontaneous variation, sociolinguistic aspects and prosody of real dialogue, limiting model performance. To overcome this, we present the *Faroese Parliament Speech Corpus (FPSC)*—a 1,600-hour collection of parliamentary recordings comprising 89,000 speeches with detailed speaker and linguistic metadata. The corpus includes weakly supervised transcriptions generated using an ensemble of four Faroese-adapted ASR models combined through a ROVER-based voting procedure. In creating *FPSC*, we trained several new state-of-the-art ASR models for Faroese—some built on large-scale pretrained backbones and others leveraging multilingual transfer—all outperforming previously published Faroese ASR systems. *FPSC* represents the first corpus of natural spoken Faroese and a major step toward realistic ASR modeling for Faroese, offering an open, reproducible, and scalable resource for future speech and language research.

Keywords: Faroese, Parliamentary Speech, ASR, Weakly-Supervised Transcription, Wav2Vec 2.0, Whisper, Metadata

1. Introduction

Automatic speech recognition (ASR) has advanced rapidly in recent years, largely driven by the availability of extensive transcribed corpora for high-resource languages. However, this progress has not been evenly distributed and many small or under-represented languages still lack the large-scale, realistic speech data required for robust model training and evaluation. Parliamentary recordings have proven to be a valuable resource, offering long-form, topic-rich speech with consistent recording conditions and clear speaker identities. They have therefore been widely used to build open ASR resources.

For Faroese, a North Germanic language spoken by around 70,000 people, such resources have been largely absent. The only publicly available dataset, the Ravnursson corpus Hernández Mena and Simonsen (2022), contains approximately 100 hours of read speech—valuable for model training but lacking the spontaneity, variability, and natural rhythm of real speech, as well as the features that distinguish spoken from written Faroese. As a result, existing Faroese ASR systems have been limited by data quantity and by the linguistic gap between scripted and naturally spoken language. Since Faroese is the official language of the Faroe Islands, parliamentary sessions of the *Løgtingið* are regularly recorded and publicly released, providing a unique opportunity to build a large, authentic corpus of natural speech. However, these recordings have no official transcripts, making them inaccessible for supervised ASR training.

To address this gap, we present the *Faroese Parliament Speech Corpus (FPSC)* í Lág (2025), a large-scale collection of parliamentary recordings paired with automatically generated transcripts. The corpus covers 368 sessions and approximately 1,600 hours of audio, comprising more than 89,000 individual speeches. Each segment includes structured metadata describing the speaker, linguistic context, and recording details.

The main contributions of this paper are the following: first we release a large-scale corpus of Faroese parliamentary recordings enriched with detailed speaker, linguistic, and technical metadata. Second, we introduce multiple newly trained Faroese ASR models—based on Wav2Vec 2.0 and Whisper—which outperform all prior Faroese ASR systems. Third, we generate transcripts for the entire corpus using four ASR models. Finally, we combine all the transcripts using a ROVER-based voting procedure (Fiscus, 1997), producing an optimal consensus transcription for each speech segment.

The paper is structured as the following: In Section 2 we give an overview of previous work, in Section 3 we discuss the structure of the *FPSC* dataset. In Section 4 we give an overview of the models used for weak transcription and their training details. In Section 5 we present statistical results on the ROVER voting scheme for producing transcripts. Finally, we discuss strengths and limita-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6196–6205
11–16 May 2026.

tions of this approach in Section 6, address ethical considerations and broader limitations in Section 7, and conclude with future work in Section 8.

2. Related Work

Several national parliaments have served as valuable resources for the development of large-scale ASR datasets Wang et al. (2021); Virkkunen et al. (2022); Solberg and Ortiz (2022); Ljubešić et al. (2022); Božík and Šuppa (2025); Helgadóttir et al. (2017, 2019); Masuyama et al. (2024); Kirkedal et al. (2020). These resources provide well-structured speech and official transcriptions, facilitating robust ASR modeling, longitudinal analysis, and cross-lingual research.

The most prominent multilingual parliamentary dataset is *VoxPopuli*, derived from European Parliament sessions and featuring over 400,000 hours of audio and nearly 2,000 hours of aligned multilingual transcriptions, enabling both supervised and semi-supervised training for a range of European languages Wang et al. (2021). Comparable initiatives have focused on single-language datasets: the *Finnish Parliament ASR Corpus* offers over 3,000 hours of transcribed speech and detailed speaker metadata, supporting research on dialectal variation and domain adaptation in ASR Virkkunen et al. (2022). Norway's *Stortinget Corpus* similarly emphasizes dialectal diversity and the democratizing effects of parliamentary data for Norwegian ASR development Solberg and Ortiz (2022).

Work on under-represented and minority languages has also advanced, but with more modest data sizes. Efforts such as the *Icelandic Althingi Corpus* have combined automatic transcription with manual post-editing to yield several hundred hours of ASR training data for a low-resource, morphologically complex language (Helgadóttir et al., 2017). Similar practices and challenges are seen in datasets for Basque, Danish, and Japanese (Masuyama et al., 2024; Kirkedal et al., 2020; Varona et al., 2024).

Weakly supervised strategies have been successfully applied in *Europarl-ASR* Garcés Díaz-Munío et al. (2021), *ParlaSpeech* Ljubešić et al. (2022), and the *Basque Parliament Corpus* Varona et al. (2024), where automatically generated transcripts served as the foundation for large-scale speech–text alignment and later refinement.

For Faroese, however, resources remain extremely limited. To date, the only publicly available ASR dataset is the *Ravnursson* corpus Hernández Mena and Simonsen (2022), comprising 100 hours speech for training and an additional 4.5 hours for validation and test splits. With it the first Wav2Vec 2.0 model for Faroese was fine-tuned yielding a Word Error Rate (WER) of 7.60% on the test set (Hernández Mena et al., 2023). While highly valuable, this resource is based on read speech and therefore does not reflect the diglossic character of Faroese, in which the formal written norm—resulting from conscious efforts of linguistic planning and standardization—diverges considerably from the spontaneous and informal features of everyday speech (Jacobsen, 2021).

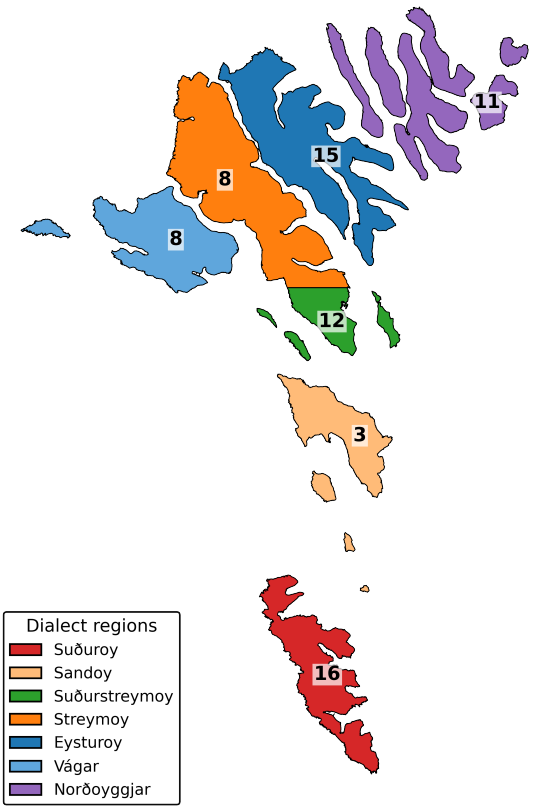

Figure 1: Geographic distribution of speakers by dialect area in the Faroe Islands

3. Data Acquisition and Processing

3.1. Source Data

The Faroese Parliament provides publicly accessible video recordings of plenary sessions on its official website[1]. For each meeting, detailed agendas are published. Altogether, 368 meetings were collected, spanning from 2020 to 2025, amounting to roughly 1,600 hours of audio and more than 89,000 individual speeches (see Table 1). The original media is stored as MP4 video streams, and parliament meetings were downloaded and converted into WAV format (16 kHz, mono).

[1] https://www.logting.fo

Description	Value
Total hours of speech	1,582
Number of meetings	368
Number of speeches	89,486
Covered years	July 2020– May 2025
Unique speakers	75
Number of dialects	7

Table 1: Summary of the *FPSC* data set

Gender	Count
Male	51
Female	23

Age group	Count
20–29	8
30–39	9
40–49	17
50–59	26
60–69	12
70–79	3

Table 2: Speaker demographics (gender and age groups).

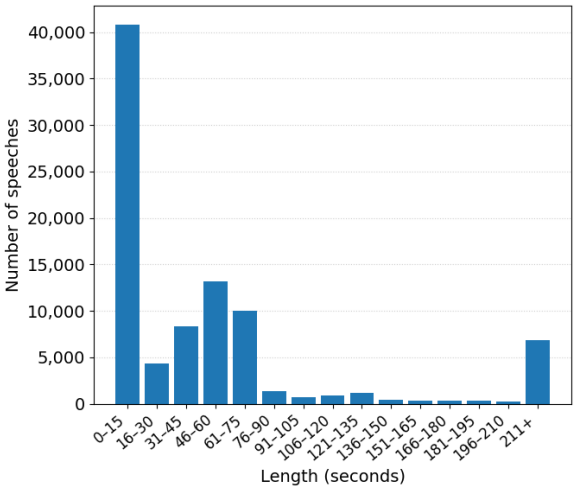

Figure 2: Speech length distribution (15 second bins; last bin = 211+ seconds)

3.2. Metadata Collection and Enrichment

Metadata was harvested directly from the parliament's meeting pages, which provide speaker names, session times, dates, topics, speech type, and speaking order. To enrich this, links to each Member of Parliament's (MP) profile page were followed to gather additional attributes such as gender and age group. See Table 2 for gender and age-group distribution. Dialect regions were inferred from place of birth and residence. See Figure 1 for an overview of dialect regions and number of MP's per region.

3.3. Speech Segmentation and Alignment

For each meeting, the parliament website provides an ordered list of speakers and their start times. We used these timestamps to align the meeting recordings with individual contributions.

In practice, about 16% of meetings contained systematic errors in the provided offsets, resulting in mismatches between audio segments and the listed speakers. These offsets were manually identified and corrected, and the metadata updated to ensure accurate alignment.

To illustrate the resulting speech units, Figure 2 presents the distribution of segment durations in 15-second bins, showing the overall variation in speech length across the corpus. Some of the speeches had time stamps with zero or negative gap and were removed.

3.4. Data Quality Issues and Corrections

Several pre-processing steps were applied to improve audio quality. Recordings were amplitude-normalized to correct for large variations in recording volume. Other metadata-related corrections included normalizing inconsistent speaker name formatting and standardizing date and time formats.

3.5. Metadata Schema

Each speech in the corpus is represented as a structured object that combines meeting-level context, speaker attributes, and technical audio properties. The schema is summarized in Table 3. This design ensures that every segment can be precisely linked back to its source, while also supporting demographic and linguistic analyses.

The meeting-related fields (`meeting_id`, `url`, `date`, `time`, `topic`, `location`) situate each speech in its institutional context. Speaker-related fields (`mp_id`, `age_group`, `dialect`, `city`, `gender`) capture biographical information, enabling analyses across demographic lines. Finally, technical descriptors (`audio_id`, `audio_format`, `sampling_rate`, `length`, `second`) provide reproducible detail about the audio files and their segmentation.

3.6. Lexical characteristics of the corpus

As part of assessing lexical coverage across different evaluation resources, we compare the most frequent unique words in the *FO-Parl-Eval-2h10m* í Lag (2025), a 2-hour and 10-minute balanced subset of *FPSC* (described in a separate publication currently under review), with the *Ravnursson* test set (Table 4). The purpose of this comparison is to illustrate the distinct lexical characteristics of a general-purpose corpus versus one drawn from parliamentary proceedings.

The *Ravnursson* test set, designed as a broad-coverage resource, shows a predominance of common verbs, adjectives, and function words (e.g.,

Field	Description
id	Unique speech segment ID
meeting_id	Meeting identifier
url	Link to meeting page
date	Meeting date (yyyy-mm-dd)
time	Speech start clock time
second	Start offset (s) within meeting
contribution_type	Role (speaker, remark, chair)
topic	Agenda item / topic
language	Language spoken (Faroese)
location	Venue (Løgtingið)
audio_format	Audio format (wav)
sampling_rate	Sample rate (16kHz)
mp_id	MP identifier
age_group	Age group (20-29, ..., 70-79)
dialect	Dialect region
city	Home city
gender	Gender
length	Duration (seconds)
audio_id	Audio filename

Table 3: Metadata fields for each speech segment.

Ravnursson Test		*FO-Parl-Eval-2h10m*	
Count	**Word**	**Count**	**Word**
25	teg	101	frú
24	eta	101	forkvinna
17	best	88	formaður
16	spurdi	43	altso
16	dámar	43	viðmerking
14	fert	39	johannesen
12	marka	29	jú
11	strika	25	joensen
11	flogfarið	22	holm
11	hennara	21	hvis
11	endurskin	20	poulsen
10	góður	20	abrahamsen
10	maðurin	19	vang
10	ilt	18	bjarni
10	tín	16	samuelsen

Table 4: Top 15 most frequent unique words in the *Ravnursson* test split and *FO-Parl-Eval-2h10m* evaluation datasets.

teg ("you"), *eta* ("eat"), *spurdi* ("asked")). By contrast, the *FO-Parl-Eval-2h10m* dataset is derived from parliamentary sessions and therefore contains a high proportion of domain-specific lexical items. Many of the most frequent unique words are named entities such as surnames of MPs (e.g., *Johannesen*, *Poulsen*, *Samuelsen*) or references to institutional roles and committees (e.g., *forkvinna* ("chairwoman"), *formaður* ("chairman"), *fíggjarnevndini* ("finance committee")). The absence of general words in the *FO-Parl-Eval-2h10m* dataset compared to *Ravnursson* reflects not only differences in domain coverage and usage patterns between simple, scripted Faroese text and parliamentary speech, but also the socio-cultural context in which they are produced—scripted, read speech versus informal everyday communication shaped by social relations and community norms.

4. Selections of models for weak transcription

To generate the weakly supervised transcripts, four Faroese-adapted ASR systems were developed and selected to maximize architectural and training diversity, balancing complementary strengths across models to enhance stability and reduce bias in the ROVER-based ensemble voting procedure. For Wav2Vec 2.0, one model was fine-tuned directly on Faroese data (í Lág, 2025c), whereas the other was continually pretrained on Faroese speech before fine-tuning (í Lág, 2025b,a). For Whisper, one model was fine-tuned on Faroese only (Hernandez Mena, 2023), and another on a multilingual mix including Norwegian and Icelandic (í Lág, 2025e), the closest linguistic and phonetic relatives of Faroese.

4.1. Continual Pretraining of Wav2Vec 2.0

To further adapt the Wav2Vec 2.0 architecture to the Faroese acoustic domain, one of the two Wav2Vec 2.0 models was fine-tuned on an existing continually pretrained model that contained 1,000 hours of parliamentary recordings from the *FPSC*. This continual pretraining phase exposed the model to a wide range of spontaneous, domain-specific speech characteristics, including variations in prosody and accents found in the data set. By fine-tuning on this adapted backbone, the model was able to leverage in-domain acoustic and linguistic knowledge, yielding stronger alignment with Faroese phonetic structures and more robust recognition performance across dialectal and demographic variation.

4.2. Multilingual Fine-Tuning of Whisper

For the first time a multilingual fine-tuning of the Whisper architecture with Faroese data has been performed. Starting from an existing Norwegian model fine-tuned on approximately 66,000 hours of parliamentary and broadcast speech (NB AI-Lab, 2024), we continued fine-tuning the model on 140 hours of Icelandic data Mollberg et al. (2020), followed by 100 hours of Faroese speech from the Ravnursson corpus Hernández Mena and Simonsen (2022).

Each stage of fine-tuning yielded consistent improvements in recognition accuracy on the Faroese evaluation set (Figure 3). The Faroese-only fine-tuned model has a WER of 4.79 (í Lág, 2025d). This was further reduced to 4.07 when the model

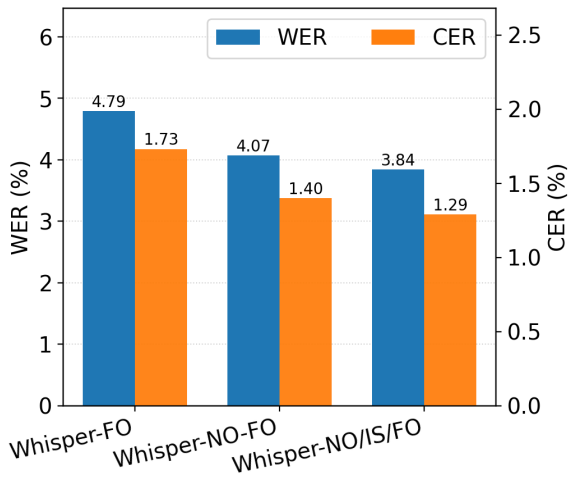

Figure 3: Multilingual continued fine-tuning of Whisper on Norwegian followed by Icelandic followed by Faroese

	FO-Parl		*Ravnursson*	
Model	**WER**	**CER**	**WER**	**CER**
Wav2Vec2-FO-CPT	26.77	**10.44**	6.11	1.75
Wav2Vec2-FO	34.07	16.69	6.79	1.95
Whisper-FO	32.56	20.24	6.58	2.19
Whisper-NO/IS/FO	**18.59**	12.19	**3.84**	**1.29**

Table 5: Performance of Wav2Vec2 and Whisper models on the *FO-Parl-Eval-2h10m* and *Ravnursson* test sets.

was fine-tuned jointly on Norwegian and Faroese data (í Lág, 2025f), and further improved to 3.84 after incorporating Icelandic (í Lág, 2025e).

4.3. Evaluation of models

To obtain transcriptions of the Faroese parliamentary speeches, we perform inference using the four selected ASR models. Specifically, we make use of the *FO-Parl-Eval-2h10m* dataset. Table 5 presents the performance of the four models on this data set, together with their results on the *Ravnursson* test set for comparison. Figure 4 shows the result for each model in terms of percentage errors on a word and character level and includes substitutions, deletions and insertions.

The results demonstrate clear differences between models at both the word and character level. The multilingual *Whisper-NO/IS/FO* model achieves the lowest overall WER, with relatively few substitutions and insertions but a higher proportion of deletions. In contrast, the *Wav2Vec2-FO-CPT* model achieves the lowest Character Error Rate (CER), suggesting more effective handling of fine-grained phonetic details and sub-word units, even if word-level segmentation remains less accurate. Among the two lower-performing models trained only on *Ravnursson*, the *Wav2Vec2-FO* model achieves a lower CER, whereas the *Whisper-FO* model performs better in terms of WER. These results highlight that model performance differ depending on whether evaluation is performed at the word or character level, reflecting complementary strengths of the different architectures and training strategies, therefore strengthening our assumption that these models provide a good base for ensemble voting in weak-supervision.

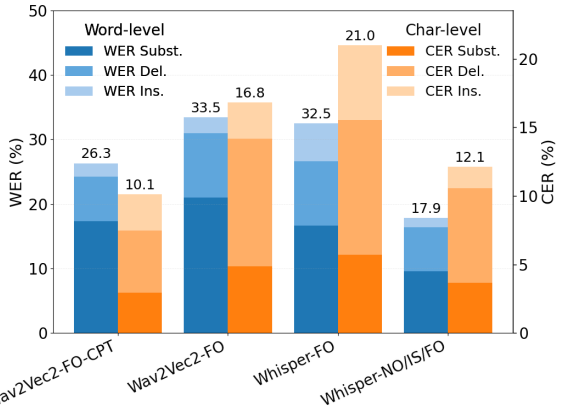

Figure 4: WER and CER per Faroese ASR model on the *FO-Parl-Eval-2h10m* data set. Bars are stacked to show the relative contribution of substitutions, deletions, and insertions.

Table 6 presents the performance of the four models evaluated on the *FO-Parl-Eval-2h10m* dataset reported by dialect, age group, and gender. This provides an insight of model consistency across demographic and linguistic factors. The four models show clear variation in performance. Again, the multilingual *Whisper-NO/IS/FO* model achieves the lowest WER across all categories, while the continually pre-trained *Wav2Vec2-FO-CPT* model attains the lowest CER values overall. Among dialects, WER values range from approximately 14% to 40%, while CER spans roughly 8% to 24% across all models. Across age groups, WER varies between about 16% and 51%, and CER between 9% and 35%. The gender split shows consistently lower error rates for female speakers across all systems.

5. Weakly-Supervised Transcription

Because no official verbatim transcripts exist, the entire *FPSC* archive was transcribed automatically using the four Faroese adopted ASR systems. Each model generated transcripts which also contain segment durations, word-level time-stamps, number of words, and other metadata that can be used for analysis.

Group	Wav2Vec2-FO		Wav2Vec2-FO-CPT		Whisper-FO		Whisper-NO/IS/FO	
	WER	CER	WER	CER	WER	CER	WER	CER
Dialect								
Eysturoy	32.3	15.5	24.5	9.8	33.7	22.4	15.1	9.4
Norðoy	36.8	18.3	29.1	10.6	39.8	23.9	21.8	14.7
Sandoy	36.7	18.4	29.7	12.9	33.7	21.9	24.0	17.7
Streymoy	34.2	16.3	28.8	10.4	28.1	13.0	18.9	16.2
Suðuroy	34.4	16.9	26.5	10.1	28.3	18.7	18.3	11.5
Suðurstreymoy	34.9	17.7	27.6	11.4	37.4	24.2	18.7	12.0
Vágar	28.1	13.3	22.5	**8.4**	24.4	14.0	**14.0**	9.1
Age Group								
20–29	34.5	16.3	28.0	10.4	32.8	19.5	18.4	10.9
30–39	29.9	14.8	23.5	9.5	25.6	15.7	**15.5**	10.5
40–49	37.0	17.8	29.2	11.1	51.0	34.5	20.4	13.9
50–59	33.0	16.1	25.4	**9.7**	29.5	18.5	17.3	11.5
60–69	34.8	18.3	28.6	11.0	34.7	21.3	19.5	12.3
70–79	38.3	18.7	29.6	12.8	25.7	15.3	20.4	11.7
Gender								
Female	29.6	14.4	23.5	**8.8**	26.9	17.7	**14.7**	9.7
Male	36.1	17.9	28.2	11.1	36.2	22.3	20.0	13.0

Table 6: WER and CER (%) per Faroese adopted ASR model, split by dialect, age group, and gender using the *FO-Parl-Eval-2h10m* data set

To obtain the most accurate representation of the spoken content, all available model transcriptions were processed through a ROVER-based voting procedure. This method aligns the transcript of multiple systems and performs a weighted majority vote at each word position. When at least two models produced identical normalized transcripts for a given segment, the transcript was accepted with full confidence; otherwise, a weighted-medoid decision was applied to resolve discrepancies. This approach reduces the expected error rate by exploiting complementary model strengths and mitigating individual system weaknesses. The final output is a single consensus transcript per speech segment that include confidence scores and detailed model rankings.

In the ROVER voting process, model outputs were weighted according to their WER and CER performance on the *FO-Parl-Eval-2h10m* evaluation data set. Specifically, we determine the weights as follows:

1. Convert error to accuracy:
   $\text{Accuracy}_{\text{word}} = 100\% - \text{WER}$
   $\text{Accuracy}_{\text{char}} = 100\% - \text{CER}$

2. Calculate combined score (S):

$$S_{\text{model}} = \frac{\text{Accuracy}_{\text{word}} + \text{Accuracy}_{\text{char}}}{2} \quad (1)$$

Model	Weight	Win Share (%)
Wav2Vec2-FO-CPT	1.04	62.10
Whisper-NO/IS/FO	1.08	18.90
Wav2Vec2-FO	0.95	10.20
Whisper-FO	0.93	8.80

Table 7: Weight and contribution of each model to the ROVER-voted transcriptions from the *FPSC* data set

3. Derive final weight (W):

$$W_{\text{model}} = \frac{S_{\text{model}}}{\frac{1}{N}\sum_{i=1}^{N} S_i} \quad (2)$$

First, the error rates for each model were converted into accuracy scores. Next, a unified performance score (S) was calculated by averaging the word and character accuracies, giving equal importance to both. Finally, these scores were normalized against the average performance of all models to derive the final weights (W). The resulting performance hierarchy is summarized in Table 7.

Across all voting decisions, 7.1% of speeches showed perfect agreement between at least two systems, while 92.9% required medoid selection. The average confidence of *0.66* indicates that most models agreed on roughly two-thirds of the words, with a small median margin of *0.056* between the winning and runner-up word-level hypotheses.

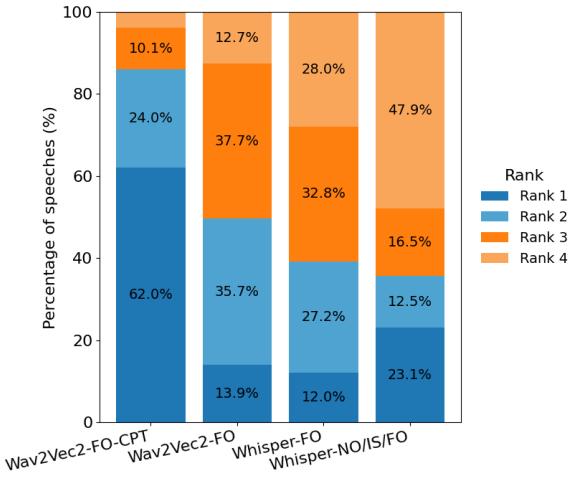

Figure 5: Performance ranking distribution of the four ASR models based on the ROVER voting scheme

Model contributions, measured as the share of winning hypotheses in the consensus transcripts, are shown in Table 7.

The continually pretrained *Wav2Vec2-CPT-FO* model contributed the majority of the winning votes (62%), followed by the multilingual *Whisper-NO/IS/FO* (19%), then *Wav2Vec2-FO* (10%), and finally *Whisper-FO* (9%). Figure 5 shows the rank distribution across the four systems. *Wav2Vec2-FO-CPT* remains the most consistent top performer, while *Whisper-NO/IS/FO* ranks second but also appears frequently in fourth place, reflecting a broader variance in its predictions. The other two models dominate the middle ranks, possibly indicating complementary behavior. Overall, these patterns confirm that continual pretraining provides stable in-domain accuracy, whereas multilingual fine-tuning introduces lexical and stylistic diversity that benefits the ROVER consensus. All transcripts are publicly available on Github [2].

6. Discussion

6.1. Establishing a new standard for Faroese ASR

The *Faroese Parliament Speech Corpus* marks a major step for Faroese ASR, enabling large-scale pretraining and realistic modeling of spontaneous speech. The corpus can benefit future systems across training paradigms—both self-supervised pretraining and multilingual fine-tuning. Although not fully error-free, the transcripts produced by the ROVER voting process provide valuable weakly supervised data that can be used to fine-tune new ASR systems. In doing so, it marks a clear advance beyond the current state of the art.

Our pipeline for corpus creation shows that a community with limited linguistic infrastructure can achieve the same methodological standards as established European corpora through open, automated, and reproducible processing. Equally important is the corpus's metadata-centric design. Each speech is embedded in detailed sociolinguistic context, making it possible to analyze dialectal and demographic variation and study the dynamics of parliamentary speech, aspects not previously accessible for Faroese. Finally, the adoption of a weakly supervised ensemble transcription pipeline illustrates how small languages can effectively scale speech resources. While the approach inevitably introduces transcription noise, it establishes a sustainable workflow where quality improves continuously as ASR systems mature.

6.2. Continual pretraining and Multilingual transfer

The comparison of ASR systems highlights the importance of two training strategies: continual pretraining and linguistic transfer from closely-related languages. Both approaches improve model performance, with one infusing specific linguistic knowledge and the other leveraging shared phonetic and prosodic structures, together yielding the best reported recognition accuracy for Faroese to date. Interestingly, the models exhibit different strengths, with continuous pretraining of Wav2Vec 2.0 obtaining better performance in terms of CER, as well as performing best in the ROVER voting system, and continuously fine-tuning of Whisper with multiple closely related languages to Faroese, obtaining lowest WER overall. Continual pretraining on large volumes of in-domain Faroese strengthened segmental accuracy; this is unsurprising as there is no real substitute for large volumes of high quality monolingual data for acquiring acoustic and linguistic knowledge. Moreover, the model was pretrained on the same parliamentary corpus, giving it a strong performance in this domain. Including closely related languages highlights the strategic value of building regional multilingual models, that are linguistically informed and take into account sociolinguistic aspects. Spoken Faroese presents a heavy usage of Scandinavian loan words, making the injection of other Scandinavian languages beneficial for reflecting real world language usage.

6.3. Weak supervision and data quality

The *FPSC* corpus demonstrates that large-scale speech resources for small languages can be produced through weak supervision when manual annotation is not an option. By aggregating outputs from several independently trained ASR systems

[2] https://github.com/davidilag/FPSC

and weighting their agreement through consensus voting, transcription quality emerges from model complementarity. This ensemble approach stabilizes predictions across architectures, yielding more consistent transcripts than any single model alone. The dominance of the continually pretrained *Wav2Vec2-FO-CPT* model highlights the strength of sustained monolingual, in-domain adaptation.

While weak supervision inevitably introduces residual errors, these are measurable and open to iterative correction. Measuring uncertainty turns residual errors into actionable feedback, enabling the corpus to improve through ongoing alignment and model refinement. In this sense, *FPSC* represents a sustainable paradigm for low-resource language technology—where scalability, openness, and continuous refinement replace the ideal of fixed completeness.

6.4. Corpus sustainability and further development

Beyond its direct contribution to Faroese ASR, the *FPSC* illustrates how public institutions can play a sustained role in the digital documentation of national languages. Parliamentary recordings, produced continuously and under consistent conditions, provide a continuous source of new recordings that combine long-term coverage with institutional credibility. By harnessing these materials within an open and reproducible framework, it is possible to establish a model for long-term language resource maintenance in small-language contexts.

All processing scripts and transcriptions from this work are released openly to promote transparency, reproducibility, and collaborative development.[3]. The corpus can be extended with new linguistic annotations, the ASR pipeline refined, or future multilingual models benchmark under identical conditions. In this way, the *FPSC* is not a static dataset but a living infrastructure designed to evolve with advances in speech technology and linguistic inquiry.

7. Ethical Considerations and Limitations

FPSC leverages parliamentary video and metadata that are freely and publicly available from the official parliament website. Accordingly, the source material falls under the public domain, in line with Faroese public records laws and established practice for legislative transparency. This enables research and downstream tasks without legal or consent-related restrictions, a framework also adopted by related parliamentary corpus projects in Finland, Norway, and Denmark Virkkunen et al. (2022); Solberg and Ortiz (2022); Kirkedal et al. (2020).

Nevertheless, several ethical and practical limitations must be acknowledged. Each speech segment in the *FPSC* is directly attributable to named parliament members, accompanied by demographic metadata; while this reflects the public nature of parliamentary proceedings, users must take care to avoid re-contextualization or profiling beyond the intended scope. Importantly, ASR-generated transcriptions may misrepresent statements, making the corpus unsuitable for verbatim or legal use. Despite efforts to represent speaker diversity, bias exists due to parliament composition and metadata extraction limits, with ASR methodologies potentially amplifying disparity among underrepresented dialects or minority speakers; users should remain cautious drawing broad sociolinguistic inferences.

8. Conclusion and Future Work

The development of the *Faroese Parliament Speech Corpus* provides a foundation for sustainable language technology in the Faroese context and possibly across other small-language communities. By transforming publicly available parliamentary recordings into a structured, openly reproducible resource, the corpus pushes the boundaries of Faroese speech research and can represent a blueprint for comparable initiatives in other minority languages.

Beyond its immediate applications in ASR, the corpus provides a foundation for longitudinal studies of how Faroese speech evolves over time and for sociolinguistic analyses examining how factors such as dialect, age, and speaker role influence language use and ASR performance. Future work will build on this foundation by further leveraging the rich metadata, using the weakly supervised transcriptions to fine-tune higher-performing ASR models, and exploiting word- and time-level alignments for improved modeling. Additional extensions—such as integrating sociopolitical and acoustic annotations or incorporating manual transcript revision through feedback loops—will deepen the corpus's analytical potential and help capture conversational dynamics, including turn-taking and interruptions.

Ultimately, the corpus exemplifies how open source data, and transparent corpus design can bridge the gap between low-resource status and linguistic innovation, ensuring that Faroese remains both technologically accessible and scientifically visible in the multilingual digital world.

[3] https://github.com/davidilag/FPSC

9. Bibliographical References

J.G. Fiscus. 1997. A post-processing system to yield reduced word error rates: Recognizer output voting error reduction (rover). In *1997 IEEE Workshop on Automatic Speech Recognition and Understanding Proceedings*, pages 347–354.

Carlos Hernández Mena, Annika Simonsen, and Jon Gudnason. 2023. ASR language resources for Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 32–41, Tórshavn, Faroe Islands. University of Tartu Library.

Carlos Daniel Hernandez Mena. 2023. Acoustic model in faroese: whisper-large-faroese-8k-steps-100h.

Jógvan í Lon Jacobsen. 2021. *Føroysk purisma: Føroysk orð ella orð í føroyskum?* Fróðskapur, Faroe University Press, Tórshavn. Published by Fróðskapur, the University of the Faroe Islands; ISBN confirmed via Google Books and PURE.

NB AI-Lab. 2024. NB-Whisper Large: A large-scale automatic speech recognition model for norwegian. Hugging Face Model Repository.

Dávid í Lág. 2025a. Wav2vec2-xls-r-300m continual pretraining on 1,000 hours of faroese parliamentary speech. https://huggingface.co/davidilag/Wav2Vec2-CPT-1000h-Fareose. Wav2Vec2-XLS-R-300M model continually pretrained on 1,000 hours of Faroese parliamentary speech without downstream ASR fine-tuning. Model hosted on Hugging Face.

Dávid í Lág. 2025b. Wav2vec2-xls-r-300m continual pretraining on faroese parliamentary speech for asr. https://huggingface.co/davidilag/wav2vec2-xls-r-300m-cpt-1000h_faroese-cp_best-faroese-100h-60-epochs_run8_2025-09-24. Wav2Vec2-XLS-R-300M model continually pretrained on 1000 hours of Faroese parliamentary speech and subsequently fine-tuned on 100 hours of Faroese speech. Model hosted on Hugging Face.

Dávid í Lág. 2025c. Wav2vec2-xls-r-300m fine-tuned for faroese speech recognition. https://huggingface.co/davidilag/wav2vec2-xls-r-300m-faroese-100h-60-epochs-20250122. Fine-tuned ASR model trained on 100 hours of Faroese speech using Wav2Vec2-XLS-R-300M. Model hosted on Hugging Face.

Dávid í Lág. 2025d. Whisper large fine-tuned on faroese speech (ravnursson 100h) for automatic speech recognition. https://huggingface.co/davidilag/whisper-large-fo-100h-30k-steps. Multilingual Whisper Large model fine-tuned on 100 hours of Faroese speech (30k training steps). Model hosted on Hugging Face.

Dávid í Lág. 2025e. Whisper large fine-tuned on norwegian, icelandic, and faroese for asr. https://huggingface.co/davidilag/whisper-large-no-is-fo-100h-30k-steps. Multilingual Whisper Large model fine-tuned on Norwegian and Icelandic parliamentary speech and subsequently on 100 hours of Faroese speech (30k training steps). Model hosted on Hugging Face.

Dávid í Lág. 2025f. Whisper large fine-tuned on norwegian parliamentary speech and faroese for asr. https://huggingface.co/davidilag/whisper-large-no-fo-100h-30k-steps. Multilingual Whisper Large model fine-tuned on Norwegian parliamentary speech and subsequently on 100 hours of Faroese speech (30k training steps). Model hosted on Hugging Face.

10. Language Resource References

Erik Božík and Marek Šuppa. 2025. *SloPalSpeech: A 2,8000-Hour Slovak Speech Corpus from Parliamentary Data*.

Gonçal V. Garcés Díaz-Munío and Joan-Albert Silvestre-Cerdà and Javier Jorge and Adrià Giménez Pastor and Javier Iranzo-Sánchez and Pau Baquero-Arnal and Nahuel Roselló and Alejandro Pérez-González-de-Martos and Jorge Civera and Albert Sanchis and Alfons Juan. 2021. *Europarl-ASR: A Large Corpus of Parliamentary Debates for Streaming ASR Benchmarking and Speech Data Filtering/Verbatimization*.

Helgadóttir, Inga Rún and Kjaran, Róbert and Nikulásdóttir, Anna Björk and Guðnason, Jón. 2017. *Building an ASR Corpus Using Althingi's Parliamentary Speeches.*

Helgadóttir, Inga Rún and Nikulásdóttir, Anna Björk and Borskỳ, Michal and Fong, Judy Y and Kjaran, Róbert and Guðnason, Jón. 2019. *The Althingi ASR System.*

Hernández Mena, Carlos Daniel and Simonsen, Annika. 2022. *Ravnursson Faroese Speech and Transcripts*.

Kirkedal, Andreas and Stepanović, Marija and Plank, Barbara. 2020. *FT Speech: Danish Parliament Speech Corpus*. ISCA, interspeech_2020.

Ljubešić, Nikola and Koržinek, Danijel and Rupnik, Peter and Jazbec, Ivo-Pavao. 2022. *ParlaSpeech-HR - a Freely Available ASR Dataset for Croatian Bootstrapped from the ParlaMint Corpus*. European Language Resources Association.

Masuyama, Mikitaka and Kawahara, Tatsuya and Matsuda, Kenjiro. 2024. *Video Retrieval System Using Automatic Speech Recognition for the Japanese Diet*. ELRA and ICCL.

Mollberg, David Erik and Jónsson, Ólafur Helgi and Þorsteinsdóttir, Sunneva and Steingrímsson, Steinþór and Magnúsdóttir, Eydís Huld and Gudnason, Jon. 2020. *Samrómur: Crowd-sourcing Data Collection for Icelandic Speech Recognition*. European Language Resources Association.

Solberg, Per Erik and Ortiz, Pablo. 2022. *The Norwegian Parliamentary Speech Corpus*. European Language Resources Association.

Varona, Amparo and Penagarikano, Mikel and Bordel, Germán and Rodriguez-Fuentes, Luis Javier. 2024. *A Bilingual Basque–Spanish Dataset of Parliamentary Sessions for the Development and Evaluation of Speech Technology*.

Anja Virkkunen and Aku Rouhe and Nhan Phan and Mikko Kurimo. 2022. *Finnish Parliament ASR corpus - Analysis, benchmarks and statistics*.

Changhan Wang and Morgane Rivière and Ann Lee and Anne Wu and Chaitanya Talnikar and Daniel Haziza and Mary Williamson and Juan Miguel Pino and Emmanuel Dupoux. 2021. *VoxPopuli: A Large-Scale Multilingual Speech Corpus for Representation Learning, Semi-Supervised Learning and Interpretation*.

í Lag, Dávid. 2025. *FO-Parl-Eval-2h10m: Faroese Parliamentary Speech Evaluation Dataset*. University of the Faroe Islands. 2h10m manually transcribed Faroese parliamentary speech evaluation dataset for ASR benchmarking.

Dávid í Lág. 2025. FPSC: Faroese parliament speech corpus. Public dataset of Faroese parliamentary speech with metadata and weakly supervised transcripts.