Stage-Aware Cross-Lingual Transfer for Faroese ASR: When and Which Languages Matter

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

Automatic speech recognition (ASR) for low-resource languages remains challenging due to limited labeled data. Although multilingual models and the inclusion of related auxiliary languages enable cross-lingual transfer, it is still unclear how introducing cross-lingual information at different training stages-pre-training versus fine-tuning-affects downstream performance. Prior work largely treats transfer as a single-stage optimization problem without disentangling stage effects. We present a stage-aware analysis of cross-lingual transfer for Faroese ASR using related auxiliary languages and Wav2Vec 2.0 XLS-R models. We systematically compare two complementary adaptation pipelines: (i) cross-lingual supervised fine-tuning and (ii) cross-lingual continuous pre-training prior to fine-tuning. Both strategies are evaluated under a unified setup with controlled model architectures, balanced representation of auxiliary languages, and identical evaluation protocols. Results demonstrate that cross-lingual transfer is stage-dependent. Supervised adaptation optimizes in-domain accuracy, while pretraining-level adaptation enhances robustness and reduces Character Error Rate (CER). Auxiliary language effects vary across pipelines, reinforcing the idea that transfer effectiveness depends on when and how cross-lingual information is introduced. Comparisons with large-scale multilingual ASR models highlight trade-offs between model scale and explicit, small-scale domain-aware adaptation. These findings suggest that effective cross-lingual transfer for Faroese low-resource ASR is inherently stage-dependent rather than a single-step design choice.

Keywords: Stage-aware cross-lingual transfer, Low-resource, Domain adaptation, Faroese language technology

1. Introduction

ASR for low-resource languages remains challenging due to limited annotated data, domain mismatch between training and deployment, and constraints on large-scale supervised learning (Baevski et al., 2020b; Conneau et al., 2020; Yadav and Sitaram, 2022; Geng et al., 2025; Müller et al., 2024). Faroese, a low-resource Scandinavian language and spoken by approximately 70,000 speakers, exemplifies these challenges: despite recent efforts Hernández Mena and Simonsen (2022) to create high-quality labeled resources, available data remain insufficient to train ASR systems that generalize beyond clean, read-speech conditions.

Recent advances in multilingual and self-supervised speech models have lowered the barrier for low-resource ASR by enabling cross-lingual transfer from high-resource or related languages (Baevski et al., 2020b; Babu et al., 2021; Radford et al., 2022b; Yadav and Sitaram, 2022). Two broad adaptation strategies are commonly used. Supervised fine-tuning directly adapts a pre-trained model to the ASR objective on the target language, optionally incorporating related languages, whereas continuous self-supervised pre-training reshapes acoustic representations prior to supervised fine-tuning. Although both strategies are well established, they are rarely compared within a unified framework for low-resource languages (Baevski et al., 2020b; Conneau et al., 2020; Babu et al., 2021; Yi et al., 2021).

Most existing work evaluates either supervised multilingual fine-tuning (Cho et al., 2018; Gupta and Boulianne, 2022; Bekarystankyzy et al., 2024; Yi et al., 2021; Williams et al., 2023) or continuous self-supervised pre-training and acoustic adaptation (Baevski et al., 2020b; Conneau et al., 2020; Babu et al., 2021; mag, 2022) in isolation, typically reporting final error rates without distinguishing between the inductive effects of cross-lingual transfer introduced during self-supervised pre-training versus supervised fine-tuning. As a result, it remains unclear whether observed gains stem from improved acoustic invariance, better phonetic alignment, or favorable domain match. This lack of stage-aware analysis forces practitioners to rely on costly trial-and-error when designing multilingual adaptation pipelines.

Moreover, auxiliary language selection adds further uncertainty to multilingual adaptation. In practice, choices are often guided by intuition or prior empirical results rather than principled criteria, making language selection and stage selection intertwined design variables. This lack of systematic guidance increases computational cost and obscures the mechanisms underlying observed trans-

Proceedings of SPEAKABLE @ LREC 2026, pages 150–161
11 May 2026.

fer gains.

To address this gap, this study investigates cross-lingual transfer strategies within the *Wav2Vec 2.0* (Baevski et al., 2020b) model family. We argue that the training stage at which cross-lingual information is introduced critically shapes transfer behavior in low-resource ASR. We therefore compare two controlled, stage-aware pipelines for Faroese ASR: (A) multilingual transfer introduced during supervised fine-tuning, and (B) multilingual transfer introduced during continuous pre-training prior to Faroese supervision. Both pipelines are evaluated against monolingual counterparts under comparable data budgets, architectures, and protocols. While both pipelines use the same underlying datasets, the allocation of data differs across training stages (self-supervised vs. supervised), reflecting their respective training strategies. Evaluation is conducted using standard Faroese test data as well as a newly constructed parliamentary speech benchmark designed to reflect real-world deployment conditions.

Our results show that the two pipelines exhibit distinct and complementary performance profiles. *Pipeline A* effectively consolidates Faroese-specific phonotactics (language-specific sound sequence constraints) and performs well on clean, in-domain speech, but remains sensitive to domain mismatch. In contrast, *Pipeline B* substantially improves robustness to acoustic variability, speaking style, and domain-specific vocabulary, yielding lower error rates on parliamentary speech. These findings suggest that cross-lingual transfer is stage-dependent, with different training stages benefiting from different types of linguistic and acoustic similarity.

To contextualize these findings, we further compare our stage-aware pipelines to two recent large-scale multilingual ASR models trained under different paradigms, using *Whisper* (Radford et al., 2022b) and *Omnilingual* (team et al., 2025) as baselines. This comparison highlights the complementary roles of size and explicit language adaptation, reinforcing the importance of training-stage and language-choice considerations in low-resource ASR.

The contributions of this paper are threefold: (1) a unified comparison of cross-lingual transfer introduced during supervised fine-tuning and during continuous self-supervised pre-training for Faroese ASR; (2) the creation of a new challenging out-of-domain Faroese parliamentary benchmark for robust evaluation; and (3) the release of newly trained Wav2Vec 2.0 models for Faroese, together with the full training and evaluation pipeline to support reproducibility and future research. While our experiments focus on Faroese, the experimental setup can serve as a blue print for low resource ASR optimization for other under-resourced languages with higher resource related languages.

The paper is organized as follows. Section 2 reviews related work. Section 3 describes the datasets and benchmark construction. Section 4 presents the stage-aware transfer pipelines. Section 5 reports empirical findings, followed by analysis in Section 6. Finally, we summarize our findings (Section 7) and discuss limitations of this study (Section 8).

2. Background and Related Work

Modern ASR systems remain highly dependent on large labeled datasets, which limits their applicability to low-resource languages. This has driven research into multilingual training, transfer learning, and self-supervised learning, where models pretrained on large amounts of unlabeled speech learn transferable representations that can be adapted with limited supervision (Besacier et al., 2014; Thomas et al., 2012). Large multilingual corpora such as Common Voice Ardila et al. (2020) have further supported this paradigm, which underlies contemporary ASR models including Wav2Vec 2.0, Whisper, and Omnilingual.

2.1. Multilingual end-to-end models: Wav2Vec 2.0, Whisper, and Omnilingual

Introduced in 2020, Wav2Vec 2.0 is a transformer-based (Vaswani et al., 2017), end-to-end self-supervised learning framework for ASR, significantly reducing reliance on large annotated data sets (Baevski et al., 2020b). Wav2Vec 2.0 employs quantized speech representations, capturing essential acoustic features across languages (Baevski et al., 2020a), facilitating multilingual learning and leveraging high-resource languages to improve ASR performance for low-resource languages (Getman et al., 2024; Williams et al., 2023). XLSR-53, a variant pretrained on 53 languages (Conneau et al., 2020), particularly demonstrates Wav2Vec 2.0's strong transfer learning capabilities through high-dimensional embeddings from large-scale multilingual training (Bekarystankyzy et al., 2024; Cho et al., 2018; Gupta and Boulianne, 2022). More recently, the XLS-R variant extended multilingual pretraining to 128 languages using a significantly larger data set (Babu et al., 2021) including 64 hours of Faroese audio from YouTube.

Whisper, released in 2022, is a multilingual ASR model trained on 680,000 hours of labeled data covering 99 languages (Radford et al., 2022a). Faroese is not included in its training data. Its encoder–decoder Transformer architecture supports both speech recognition and speech trans-

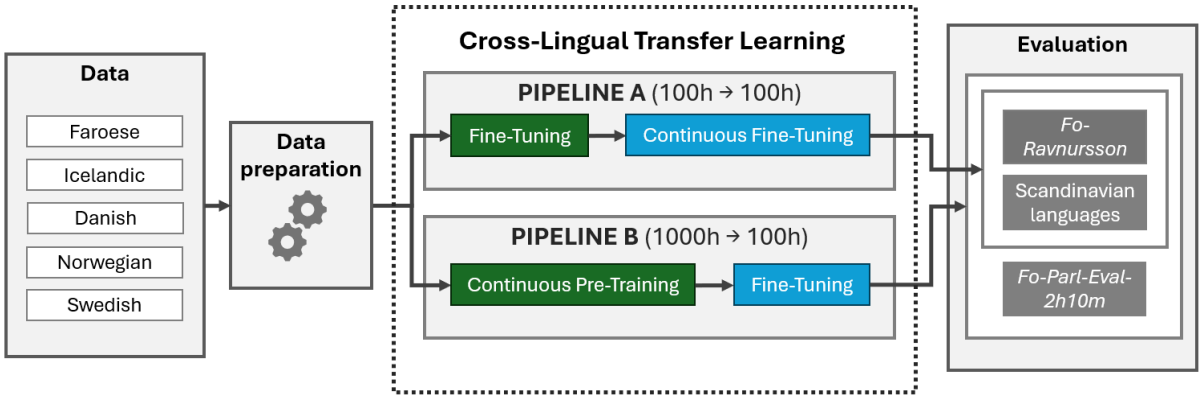

Figure 1: Stage-aware cross-lingual transfer framework for Faroese ASR. The diagram shows four components: (1) multilingual datasets, (2) unified data preparation, (3) two alternative adaptation pipelines, and (4) evaluation on in-domain and out-of-domain Faroese test sets. Pipeline A performs supervised cross-lingual fine-tuning followed by Faroese-only continued fine-tuning, whereas Pipeline B introduces cross-lingual information earlier through continuous pre-training before Faroese fine-tuning. Green blocks indicate stages where auxiliary or multilingual data are used; blue blocks denote Faroese-only supervised training.

lation. Whisper is trained end-to-end using supervised learning on diverse real-world data, improving robustness to noise, accents, and speaking styles.

Recent work has extended multilingual ASR beyond fixed language inventories through the Omnilingual ASR framework, released in November 2025 (team et al., 2025). The framework scales Wav2Vec 2.0-style self-supervised learning to models of up to 7B parameters, pretrained on over 4.3 million hours of speech spanning more than 1,600 languages (team et al., 2025).

While these models demonstrate that large-scale multilingual exposure is effective, they offer limited insights into how different languages affect each other, and when cross-lingual information is most beneficial to a specific low-resource language, therefore requiring an extra, empirical optimization step for low-resource adaptation.

2.2. ASR for Faroese

Faroese ASR development accelerated with the Ravnur project, which produced the Faroese *BLARK* and the 100-hour Ravnur corpus covering multiple dialects and age groups (Simonsen et al., 2022). Using this data, fine-tuned multilingual models such as Wav2Vec 2.0 achieved the first competitive Faroese ASR system (7.6% WER) (Hernandez Mena et al., 2023). Subsequent work introduced grapheme-to-phoneme modeling and expanded linguistic resources for standardized processing (Lamhauge et al., 2023). A recent representation-space analysis of Wav2Vec 2.0 XLSR-53 further examined Faroese in relation to 102 languages and found that Swedish and Norwegian emerge as its closest Scandinavian neighbors in different encoder layers, based on Euclidean distances in the representation space(í Lág et al., 2024).

Despite its close genealogical relationship to Icelandic, Faroese differs substantially in phonology and orthographic conventions. It exhibits rich inflectional morphology, productive compounding, vowel quantity contrasts, and complex consonant clusters (Thráinsson et al., 2012; Eghdam and co-authors incl. user, 2023; Petersen and Voeltzel, 2025). These characteristics increase lexical variability and complicate end-to-end modeling under limited supervision. In addition, extensive historical contact with Danish has resulted in lexical borrowing and code-mixing, particularly in formal domains such as parliamentary speech (Lamhauge et al., 2023; Hernandez Mena et al., 2023).

3. Data

3.1. Data sets

This study draws on a curated set of publicly available speech corpora covering Faroese and four closely related North Germanic languages (Icelandic, Danish, Norwegian, and Swedish). The data are sourced from publicly available collections hosted on the *Huggingface* platform and selected based on recent release and annotation quality.

The corpora are organized according to the two pipelines. Pipeline A uses labeled Faroese speech from *Ravnursson* and mixed-domain labeled data for the four auxiliary languages. Pipeline B uses unlabeled parliamentary speech for all languages during continuous pre-training, including Faroese *FPSC* parliament data í Lág (2025) and *Ravnurs-*

son for subsequent fine-tuning.

Data for *Pipeline A*. Faroese: *Ravnursson* Hernández Mena and Simonsen (2022), (url, 2022b); Icelandic: *Samrómur* O'Brien et al. (2024), (url, 2020); Danish: *CoRal* Nielsen et al. (2024), (url, 2024); Norwegian: *Norwegian Parliamentary Speech Corpus (NPSC)* Solberg and Ortiz (2022), (url, 2022a); Swedish: *RixVox* Rekathati (2023), (url, 2023).

Data for *Pipeline B*. Faroese: *Faroese Parliament Speech Corpus (FPSC)* í Lág (2025), using a 1000-hour subset of parliamentary debates from 2020–2025 í Lág (2025) and *Ravnursson* Hernández Mena and Simonsen (2022); Icelandic: *Althingi Parliamentary Speech* (2005–2016) Helgadóttir et al. (2021); Danish: *FT Speech – Danish Parliament Speech Corpus* (2010–2019) Kirkedal et al. (2020); Norwegian: *NST - Norwegian ASR Database*, parliamentary speech from 2017–2018 spr (2023); Swedish: *RixVox* parliamentary recordings (2003–2023) Rekathati (2023), (url, 2023).

Evaluation data: *Ravnursson* test split and the new *FO-Parl-Eval-2h10m* benchmark All models are evaluated on the official *Ravnursson* test split, corresponding to the dataset used for supervised fine-tuning. In addition, we introduce a new Faroese evaluation benchmark, *FO-Parl-Eval-2h10m* í Lag (2025), comprising 2 hours and 10 minutes of curated parliamentary speech extracted from the *Faroese Parliament Speech Corpus (FPSC)* í Lag (2025). The dataset consists of 344 speech segments ranging from 5 to 29 seconds and reflects real-world parliamentary speech conditions. It is designed to be representative in terms of speaker gender, age groups, and dialectal coverage. As parliamentary speech involves recurring speakers, the evaluation is speaker-dependent, meaning that speakers may appear in both training and evaluation data. This setup reflects realistic deployment conditions but should be interpreted accordingly when assessing generalization. To prevent data leakage at the segment level, all segments included in *FO-Parl-Eval-2h10m* are excluded from any data used in the continuous pretraining stage in Pipeline B.

3.2. Preparation of Data Sets for Experiments

Data preparation is structured around the two experimental pipelines. Data are partitioned and preprocessed differently in each pipeline depending on their role in self-supervised adaptation or supervised training, ensuring stage-appropriate use and comparability across pipelines.

Preparation of Data for Pipeline A. For each auxiliary language (Icelandic, Danish, Norwegian, and Swedish), we first collect all available audio–text data from the respective corpora. From this pool, we randomly sample speech segments until a total duration of 100 hours is reached. This results in a balanced dataset of 100 hours per language.

Formally, let D_l denote the full set of available audio–text pairs for language l, where each sample consists of an audio signal a and its transcription t. We construct a subset $D_l^{100h} \subset D_l$ such that the total duration satisfies:

$$\sum_{(a,t)\in D_l^{100h}} \text{Duration}(a) \approx 100\,\text{h}. \tag{1}$$

The four language-specific subsets are then combined to form a multilingual dataset:

$$D_{\text{multi}} = \bigcup_l D_l^{100h} \tag{2}$$

resulting in a total of 400 hours of speech. Each sample is annotated with its corresponding language identifier. Finally, the combined dataset is split into training, validation, and test partitions for supervised fine-tuning Lág (2025).

Preparation of Data for Pipeline B. From the 1,600-hour *Faroese Parliament Speech Corpus (FPSC)* í Lág (2025), 1000 hours of speech were randomly selected. Non-speech and extremely low-volume segments were removed, and all recordings were normalized in volume. The resulting data were segmented into fixed-length audio chunks of 20 seconds, yielding approximately 180,000 audio files.

For the four auxiliary Scandinavian languages, 500 hours per language were randomly sampled from their respective parliamentary speech corpora and processed using the same pipeline. To ensure consistent cross-lingual exposure during training, the data for Faroese and each auxiliary language were mixed such that consecutive speech segments alternated between Faroese and the auxiliary language. Each language was segmented into 90,000 audio files of 20 seconds. For the combined dataset with all five languages, 200 hours per language were randomly selected from the respective subsets and systematically interleaved to ensure balanced mixing across languages.

4. Methods

4.1. Experiments

An overview of the two adaptation pipelines and their shared components is shown in Figure 1. The pipelines differ in the stage at which cross-lingual information is introduced: Pipeline A incorporates it during supervised fine-tuning, whereas Pipeline B introduces it through an intermediate continuous self-supervised pre-training stage prior to fine-tuning.

Pipeline A consists of two supervised fine-tuning stages—first on the auxiliary language, and subsequently on Faroese. In contrast, Pipeline B inserts a continuous multilingual pre-training stage between the original XLS-R pre-training and the final supervised fine-tuning step. Although XLS-R is initially pretrained on 128 languages, all intermediate adaptation stages in this study operate on controlled language subsets. Specifically, these stages use either a single Scandinavian auxiliary language or a balanced mixture of Scandinavian languages, while the final supervised ASR fine-tuning is performed exclusively on Faroese.

All models are evaluated after a final stage of supervised fine-tuning on the same 100-hour labeled Faroese dataset, ensuring a consistent target-language adaptation across all configurations. Evaluation is performed on the *Ravnursson* test set and the *FO-Parl-Eval-2h10m* benchmark.

For Pipeline A, Wav2Vec 2.0 XLS-R models are evaluated in the 300M, 1B, and 2B parameter variants. Due to computational constraints, Pipeline B is evaluated only using the 1B variant. As architectural baselines, Whisper and Omnilingual ASR models are included for comparison. Whisper is directly fine-tuned on Faroese under the same target-language conditions. Omnilingual models are used solely for inference as external baselines and are not adapted. They are not evaluated on the *Ravnursson* test split due to potential overlap with their training data (Meta AI Research, 2025), and to ensure a fair comparison with models evaluated on disjoint data.

All code, configuration files, and data processing pipelines required to reproduce the experiments are publicly available on GitHub[1]

Pipeline A: Supervised Fine-Tuning–Based Adaptation Pipeline A introduces cross-lingual transfer exclusively through labeled supervision. No additional self-supervised training is performed beyond the original multilingual pretraining.

The pipeline consists of two supervised stages per of the six experiments:

1. *FT (Fine-Tuning stage):* The pretrained XLS-R model is fine-tuned on 100 hours of labeled speech from one or more auxiliary Scandinavian languages.

2. *CFT (Continued Fine-Tuning stage):* The model is further fine-tuned on 100 hours of labeled Faroese only.

We compare the results from these experiments with a monolingual baseline, where both FT and CFT stages are performed with Faroese only.

Configurations

We define six experiments (E1–E6). E1 serves as the monolingual baseline. Experiments E2–E5 each incorporate a single Scandinavian auxiliary language during the fine-tuning (FT) phase, while E6 uses a balanced mixture of Scandinavian languages during fine-tuning.

- *FT-E1 (Monolingual baseline):*
  FT(FO $= 100$h) $\rightarrow$ *CFT*(FO $= 100$h)
- *FT-E2–E5 (Single auxiliary language):*
  FT($X = 100$h, $X \in$ {IS, DK, NO, SW}) $\rightarrow$
  CFT(FO $= 100$h)
- *FT-E6 (Multilingual mixture):*
  FT(25h per X, $X \in$ {IS, DK, NO, SW}) $\rightarrow$
  CFT(FO $= 100$h)

All models are made publicly available (í Lág, 2025b,c,a,d,e,f).

Pipeline B: Continuous Self-Supervised Pre–Training–Based Adaptation Pipeline B introduces cross-lingual transfer at the representation level via continuous pre-training before supervised fine-tuning.

The pipeline consists of:

1. *CPT (Continuous Pre-Training stage):* The XLS-R model is further trained using the original Wav2Vec 2.0 self-supervised objective on 1000 hours unlabeled speech. The 1000 hours of unlabeled speech contain one or more auxiliary Scandinavian languages.

2. *FT (Fine-Tuning stage):* The model is then fine-tuned on 100 hours of labeled Faroese.

Configurations

We define six experiments (E1–E6). E1 serves as the monolingual baseline. Experiments E2–E5

[1] https://github.com/davidilag/Stage-Aware-Cross-Lingual-Transfer-for-Faroese-ASR-2026

each incorporate 500 hours from a single Scandinavian auxiliary language in the CPT phase, accompanied by 500 of Faroese. E6 uses a balanced mixture (200 hours each) of Scandinavian languages, including Faroese, in the CPT phase.

- *CPT-E1 (Monolingual baseline):* *CPT*(FO $= 1000$h) $\rightarrow$ *FT*(FO $= 100$h)
- *CPT-E2–E5 (Balanced bilingual):* *CPT*(FO $= 500$h $+ X = 500$h, $X \in$ {IS, DK, NO, SW}) $\rightarrow$ *FT*(FO $= 100$h)
- *CPT-E6 (Multilingual mixture):* *CPT*(200h per X, $X \in$ {FO, IS, DK, NO, SW}) $\rightarrow$ *FT*(FO $= 100$h)

All models are made publicly available (í Lág, 2025g,j,i,k,l,h)

4.2. Whisper and Omnilingual Models

To provide a comparison with state-of-the-art ASR systems, we include Whisper and Omnilingual as reference models. Whisper model variants 39M, 74M, 244M, 1.55B are fine-tuned exclusively on Faroese speech and evaluated on the *Ravnursson* test set. Omnilingual 7B models are already trained on *Ravnursson* and are therefore evaluated in inference-only mode without additional adaptation on *FO-Parl-Eval-2h10m*.

4.3. Formal Definition of Transfer Setups

We formalize the transfer setups using two generic adaptation operators applied to a pretrained multilingual encoder θ_0.

Let D_s^{lab} and D_s^{unlab} denote labeled and unlabeled data from one or more source languages, and D_t^{lab} and D_t^{unlab} labeled and unlabeled data from the target language (Faroese).

Adaptation operators. We define:

$$\mathcal{F}_{\text{ASR}}(\theta; D) = \arg\min_{\theta} \mathcal{L}_{\text{ASR}}(D; \theta), \tag{3}$$

$$\mathcal{P}_{\text{SSL}}(\theta; D) = \arg\min_{\theta} \mathcal{L}_{\text{SSL}}(D; \theta), \tag{4}$$

where $\mathcal{L}_{\text{ASR}}$ denotes the supervised speech recognition training objective and $\mathcal{L}_{\text{SSL}}$ denotes a self-supervised representation learning objective applied to unlabeled speech data.

Transfer setups. Using these operators, the training pipelines are expressed as:

$$\theta_{\text{FT}} = \mathcal{F}_{\text{ASR}}(\theta_0; D_t^{\text{lab}}) \tag{5}$$

$$\theta_{\text{CFT}} = \mathcal{F}_{\text{ASR}}(\theta_0; D_s^{\text{lab}} \cup D_t^{\text{lab}}) \tag{6}$$

$$\theta_{\text{CPT}} = \mathcal{P}_{\text{SSL}}(\theta_0; D_s^{\text{unlab}} \cup D_t^{\text{unlab}}) \tag{7}$$

$$\theta_{\text{CPT}\rightarrow\text{FT}} = \mathcal{F}_{\text{ASR}}\Big(\mathcal{P}_{\text{SSL}}(\theta_0; D_s^{\text{unlab}} \cup D_t^{\text{unlab}}); D_t^{\text{lab}}\Big) \tag{8}$$

Pipeline / Exp.	*Ravnursson* WER	*Ravnursson* CER	*FO-Parl* WER	*FO-Parl* CER
Pipeline A				
FT-E1 (FO)	7.49	2.13	37.78	17.70
FT-E2 (FO-IS)	8.29	2.36	42.00	20.20
FT-E3 (FO-DK)	7.75	2.22	39.39	18.31
FT-E4 (FO-NO)	**6.91**	**2.04**	36.92	**17.16**
FT-E5 (FO-SW)	7.24	2.06	36.68	17.64
FT-E6 (All)	7.38	2.13	**36.61**	17.35
Pipeline B				
CPT-E1 (FO)	**6.59**	**1.85**	**26.46**	**10.68**
CPT-E2 (FO-IS)	6.80	1.95	27.64	10.96
CPT-E3 (FO-DK)	6.96	1.97	26.90	10.84
CPT-E4 (FO-NO)	7.08	2.00	26.68	10.86
CPT-E5 (FO-SW)	7.29	2.06	31.06	14.72
CPT-E6 (All)	7.47	2.11	29.80	12.72

Table 1: Evaluation results for two training pipelines on the *Ravnursson* and *FO-Parl-Eval-2h10m* test sets. Highest performing models are listed in bold text and lowest in red text.

This formulation isolates training stage from data selection, allowing the two pipelines to be interpreted as alternative compositions of the same adaptation operators, differing only in when cross-lingual and target-specific information is introduced.

5. Results

Table 1 summarizes performance under the two pipelines on two Faroese test sets. Performance is consistently lower on *FO-Parl-Eval-2h10m* than on the *Ravnursson*, indicating that the parliamentary benchmark is substantially more challenging and better reflects difficult real-world conditions. Across all language configurations, Pipeline B yields clear gains over Pipeline A on *FO-Parl-Eval-2h10m*, with the largest absolute improvements observed for Faroese-only training.

A second observation is that the optimal auxiliary language depends on the training strategy. In Pipeline B, single-language Scandinavian auxiliaries (notably Icelandic, Danish, and Norwegian) typically outperform the balanced five-language mixture, whereas in Pipeline A, gains are smaller and less consistent, with Norwegian and Swedish providing the strongest improvements among the single-language variants.

As shown in Figure 3, continuous fine-tuning (Pipeline A) improves performance for Faroese with all auxiliary languages except Icelandic. Norwegian, as auxiliary language, yields the strongest improvement for Faroese, followed by Swedish, the Scandinavian mixture, and Danish. In contrast, Icelandic does not provide a measurable benefit over direct fine-tuning.

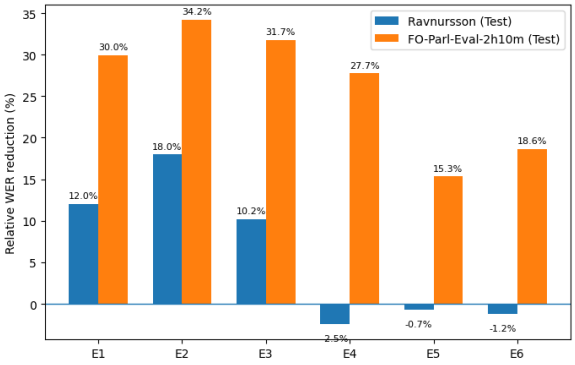

Figure 2: Relative WER reduction (%) of Pipeline B over Pipeline A by experiment (E1–E6) on *Ravnursson* and *FO-Parl-Eval-2h10m* evaluation data sets.

Finally, model capacity and training duration matter for Faroese-only fine-tuning. Table 2 shows that extending training from 30 to 60 epochs improves all XLS-R sizes, with the 1B variant providing the strongest Faroese-only baseline on *Ravnursson*. Table 3 shows clear scaling for Whisper, where larger models substantially reduce error rates and establish a competitive supervised baseline.

Figure 2 summarizes the relative error reduction achieved by Pipeline B over Pipeline A across experiments *E1–E6*. Gains are consistently larger on the parliamentary benchmark, confirming that CPT primarily improves robustness under more challenging conditions, while improvements on the *Ravnursson* test set are smaller and language dependent.

Model Size	30 epochs		60 epochs	
	WER (%)	CER (%)	WER (%)	CER (%)
XLS-R 300M	**7.46**	**2.08**	6.78	1.95
XLS-R 1B	7.49	2.13	**5.85**	**1.73**
XLS-R 2B	7.94	2.28	6.58	1.91

Table 2: Wav2Vec 2.0 XLS-R fine-tuned exclusively on Faroese for 30 and 60 epochs (FT-E1), evaluated on *Ravnursson*.

Model Size	WER (%)	CER (%)
Tiny (39M)	35.92	13.47
Base (74M)	22.86	7.40
Small (244M)	15.03	4.75
Large (1.55B)	**6.51**	**2.11**

Table 3: Whisper fine-tuned exclusively on 100h Faroes, evaluated on *Ravnursson*.

In table 4 we compare our best performing Wav2Vec 2.0 model with two Omnilingual 7B models evaluated on *FO-Parl-Eval-2h10m*. The Omnilingual LLM variant achieves the lowest WER,

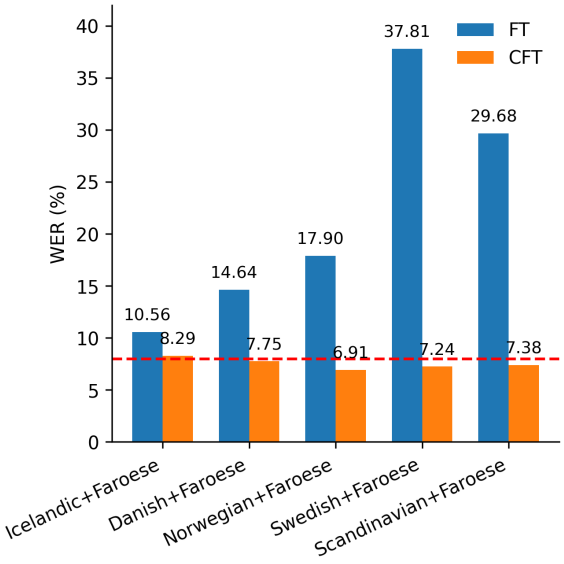

Figure 3: WER after multilingual fine-tuning and continuous fine-tuning (Pipeline A) using different Scandinavian source-language configurations. The dashed red line indicates a reference WER level (Faroese only).

Model	Variant	WER (%)	CER (%)
Wav2Vec2	CPT→FT	26.46	10.68
Omnilingual 7B	LLM	23.06	11.71
Omnilingual 7B	CTC	35.99	15.51

Table 4: Performance on *FO-Parl-Eval-2h10m* data set. Wav2Vec 2.0 model is the best perming model in the Pipeline B. Omnilingual models are evaluated in inference-only mode since it already has been fine-tuned on *Ravnursson*.

while the Pipeline B model attains comparable performance with substantially lower CER. In contrast, the Omnilingual CTC variant degrades markedly. These results show that targeted language adaptation in a continual pre-training setting can still provide domain-specific robustness over large, massively multilingual, out of the box models.

6. Discussion

Our comparative analysis demonstrates that cross-lingual transfer to Faroese behaves differently across adaptation stages and auxiliary language configurations, with distinct performance patterns observed for supervised fine-tuning and continuous pre-training. Performance does not follow a uniform pattern across languages; rather, the effectiveness of a given auxiliary language depends on whether cross-lingual exposure is introduced during continuous self-supervised pre-training (Pipeline A) or during supervised fine-tuning (Pipeline B).

Within Pipeline A, Norwegian and Swedish consistently provide the strongest gains when used as auxiliary languages, despite not exhibiting the best standalone ASR performance. We speculate the reason for this might be phonetic proximity as primary driver of transfer effectiveness during fine-tuning. Faroese and Norwegian, in particular, share a strong similarity in acoustic–phonetic structure—including consonant realization, vowel quantity, stress patterns, palatalization, and prosody—possibly facilitating parameter adaptation in later supervised fine-tuning stages, where the model specializes to Faroese phonotactics and grapheme-to-phoneme mappings. In contrast, Icelandic, although genealogically closest to Faroese, contributes limited improvement in this pipeline, suggesting that phylogenetic relatedness alone is insufficient to enhance transfer during fine-tuning. Norwegian and Swedish also proved to be the closest to Faroese within the Wav2Vec 2.0 representation space (í Lág et al., 2024), suggesting perceived proximity by the model may facilitate transfer.

In Pipeline B, transfer dynamics differ substantially. Continuous pre-training operates at the representation level and appears less sensitive to fine-grained phonetic similarity. Instead, gains reflect broader structural and lexical alignment. Icelandic and Danish yield improvements during CPT, likely due to morphosyntactic overlap and stylistic similarity in parliamentary speech. The strong effect of Danish on the parliamentary evaluation set supports this interpretation, as formal Faroese discourse contains a high density of Danish loanwords.

Taken together, the two pipelines suggest complementary mechanisms: Pipeline B enhances robustness to real world language usage, while Pipeline A sharpens phonetic and pronunciation-specific modeling. The auxiliary language that is optimal in one pipeline is not necessarily optimal in the other, reinforcing that cross-lingual transfer must be treated as a stage-aware and language-specific optimization problem.

Comparison with fine-tuned Whisper and out-of-the-box Omnilingual yields additional perspective. Whisper, fine-tuned on Faroese data, demonstrates strong word-level robustness due to large-scale supervised multilingual training. On *FO-Parl-Eval-2h10m*, the Omnilingual 7B LLM variant achieves the best result, outperforming our best Pipeline B configuration, while the same Pipeline B model attains lower CER, indicating more stable character-level modeling. These differences suggest complementary error profiles: large multilingual models improve word-level prediction, whereas continuous pre-training remains advantageous for sub-word precision and language-specific orthographic consistency.

7. Conclusions

This work demonstrates that cross-lingual transfer for low-resource Faroese ASR is inherently stage-dependent, and that the timing of cross-lingual information injection fundamentally shapes downstream behavior. Through a unified comparison of two adaptation pipelines, we show that the stage at which auxiliary languages are introduced determines whether transfer learning primarily benefits from phonetic similarity or from syntactic/lexical overlap.

A key contribution of this study is the controlled, side-by-side evaluation of both pipelines under closely matched language data representation, shared model architectures, and consistent evaluation protocols. This isolates the effect of adaptation technique and auxiliary language composition. Pipeline A seemingly consolidates Faroese-specific phonotactics and performs strongly in-domain, whereas Pipeline B consistently improves spelling (lowest CER) and robustness under parliamentary speech conditions. The two pipelines are therefore complementary rather than interchangeable.

Beyond Faroese, this work introduces new insights on stage-aware adaptation for structuring low-resource, cross-lingual ASR. By reframing multilingual transfer as a design choice over training strategies and data composition, the study contributes a methodological lens for analyzing adaptation strategies in low-resource ASR within self-supervised and large-scale multilingual settings.

8. Limitations

This study has several limitations. First, due to computational constraints, we focus on smaller model variants, and continuous pre-training (CPT) is conducted for only one model size. While this limits conclusions about scalability, it reflects realistic deployment conditions in low-resource settings, where computational resources are limited.

Second, our analysis focuses on a single target language, Faroese, with specific sociolinguistic characteristics. Although this provides a focused case study, the generalizability of our findings to other low-resource languages remains to be validated.

Finally, the continuous pre-training stage relies on data from the same domain as one evaluation set (*FO-Parl-Eval-2h10m*). This may advantage CPT-adapted models on that dataset. However, given the severe data constraints typical of low-resource languages, avoiding domain overlap is often impractical, making this limitation representative of real-world conditions rather than an isolated methodological choice.

9. Bibliographical References

2020. Samrómur icelandic speech corpus. https://huggingface.co/datasets/language-and-voice-lab/samromur_asr.

2022. Magic dust for cross-lingual adaptation of monolingual wav2vec-2.0. In *ICASSP 2022 – 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*.

2022a. Norwegian parliamentary speech corpus (npsc) by nbailab. https://huggingface.co/datasets/NbAiLab/NPSC.

2022b. Ravnursson corpus. https://huggingface.co/datasets/carlosdanielhernandezmena/ravnursson_asr.

2023. Rixvox, swedish parliament speech from period 2003-2023. https://huggingface.co/datasets/KBLab/rixvox.

2024. Coral: Danish conversational and read-aloud dataset. https://huggingface.co/datasets/alexandrainst/coral.

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. 2021. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*.

Alexei Baevski, Alexis Conneau, and Michael Auli. 2020a. Wav2vec 2.0: Learning the structure of speech from raw audio. *Meta AI*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020b. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Akbayan Bekarystankyzy, Orken Mamyrbayev, Mateus Mendes, Anar Fazylzhanova, and Muhammad Assam. 2024. Multilingual end-to-end asr for low-resource turkic languages with common alphabets. *Scientific Reports*, 14(1):13835.

Laurent Besacier, Etienne Barnard, Alexey Karpov, and Tanja Schultz. 2014. Automatic speech recognition for under-resourced languages: A survey. In *Speech Communication*, volume 56, pages 85–100.

Jaejin Cho, Murali Karthick Baskar, Ruizhi Li, Matthew Wiesner, Sri Harish Mallidi, Nelson Yalta, Martin Karafiat, Shinji Watanabe, and Takaaki Hori. 2018. Multilingual sequence-to-sequence speech recognition: architecture, transfer learning, and language modeling. In *2018 IEEE Spoken Language Technology Workshop (SLT)*, pages 521–527. IEEE.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2020. Unsupervised cross-lingual representation learning for speech recognition. *arXiv preprint arXiv:2006.13979*.

Iben Eghdam and co-authors incl. user. 2023. Standardising pronunciation for a grapheme-to-phoneme converter for faroese. In *Proceedings of the 24th Annual Conference of the International Speech Communication Association (INTERSPEECH 2023)*, page to appear. Preprint version consulted.

Yutian Geng, Hilary Dempster, Robert Jimerson, Martin Haspelmath, Luke Zettlemoyer, and Shinji Watanabe. 2025. Evaluating speech foundation models for automatic speech recognition in the low-resource kanyen'kéha language. In *Proc. Interspeech*. Despite recent progress in SFMs, low-resource Indigenous ASR remains challenging due to limited annotated data and extensive vocabulary variation.

Yaroslav Getman, Tamás Grósz, Katri Hiovain-Asikainen, and Mikko Kurimo. 2024. Exploring adaptation techniques of large speech foundation models for low-resource asr: a case study on northern sámi. *Interspeech 2024*.

Vishwa Gupta and Gilles Boulianne. 2022. Progress in multilingual speech recognition for low resource languages kurmanji kurdish, cree and inuktut. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6420–6428.

Carlos Hernandez 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.

Dávid í Lág, Barbara Scalvini, and Jon Gudnason. 2024. Mapping faroese in the multilingual representation space: Insights for ASR model optimization. In *The Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies*.

Sandra Lamhauge, Iben Debess, Carlos Hernández Mena, Annika Simonsen, and Jon Gudnason. 2023. Standardising pronunciation for a grapheme-to-phoneme converter for Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 308–317, Tórshavn, Faroe Islands. University of Tartu Library.

Meta AI Research. 2025. Omnilingual ASR: lang_ids.py. https://github.com/facebookresearch/omnilingual-asr/blob/main/src/omnilingual_asr/models/wav2vec2_llama/lang_ids.py. GitHub repository. Accessed: 2026-02-16.

Sonja Müller, Daniel Fuchs, and Laurette Pretorius. 2024. Exploring asr fine-tuning on limited domain-specific data for low-resource languages. In *Southern African Linguistics and Applied Language Studies*. Shows that modern ASR trained on government/political data performs poorly on out-of-domain broadcast news, underscoring domain mismatch under low-resource conditions.

Hjalmar P. Petersen and Laurence Voeltzel. 2025. *Faroese Phonetics and Phonology*, volume 34 of *Phonology and Phonetics*. De Gruyter Mouton, Berlin.

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

Alec Radford et al. 2022b. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*. Introduces Whisper, a multilingual model trained on 680k hours of weakly supervised data, showing strong zero-/few-shot transfer to many low-resource languages.

Annika Simonsen, Sandra Saxov Lamhauge, Iben Nyholm Debess, and Peter Juel Henrichsen. 2022. Creating a basic language resource kit for faroese. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4637–4643.

Omnilingual ASR team, Gil Keren, Artyom Kozhevnikov, Yen Meng, Christophe Ropers, Matthew Setzler, Skyler Wang, Ife Adebara, Michael Auli, Can Balioglu, Kevin Chan, Chierh Cheng, Joe Chuang, Caley Droof, Mark Duppenthaler, Paul-Ambroise Duquenne, Alexander Erben, Cynthia Gao, Gabriel Mejia Gonzalez, Kehan Lyu, Sagar Miglani, Vineel Pratap, Kaushik Ram Sadagopan, Safiyyah Saleem, Arina Turkatenko, Albert Ventayol-Boada, Zheng-Xin Yong, Yu-An Chung, Jean Maillard, Rashel Moritz, Alexandre Mourachko, Mary Williamson, and Shireen Yates. 2025. Omnilingual asr: Open-source multilingual speech recognition for 1600+ languages.

Samuel Thomas, Michael L. Seltzer, Kenneth Church, and Hynek Hermansky. 2012. Deep neural network features and semi-supervised training for low-resource speech recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6704–6708.

Höskuldur Thráinsson, Hjalmar P. Petersen, Jógvan í Lon Jacobsen, and Zakaris Svabo Hansen. 2012. *Faroese: An Overview and Reference Grammar*, 2 edition. Fróðskapur, Tórshavn.

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

A. Williams, A. Demarco, and C. Borg. 2023. The applicability of wav2vec2 and whisper for low-resource maltese asr. In *Proceedings of SIGUL 2023*.

Hemant Yadav and Sunayana Sitaram. 2022. A survey of multilingual models for automatic speech recognition. *Language Resources and Evaluation*. Surveys multilingual and cross-lingual ASR, emphasizing lack of labeled data for most languages and the role of transfer/self-supervised learning in low-resource settings.

C. Yi, J. Wang, C. Ning, S. Zhou, and B. Xu. 2021. Transfer ability of monolingual wav2vec2.0 for low-resource speech recognition. In *Proceedings of the International Joint Conference on Neural Network*, pages 1–6.

Dávid í Lág. 2025a. Wav2vec2 xls-r 1b: Fine-tuning on 100h danish followed by continuous fine-tuning on 100h faroese (ft→cft, e3). https://huggingface.co/davidilag/wav2vec2-xls-r-1b-E3-faroese-100h-30-epochs_20250124_v2.

Dávid í Lág. 2025b. Wav2vec2 xls-r 1b: Fine-tuning on 100h faroese (ft, e1). https://huggingface.co/davidilag/wav2vec2-xls-r-1b-faroese-100h-60-epochs_20250108_v2.

Dávid í Lág. 2025c. Wav2vec2 xls-r 1b: Fine-tuning on 100h icelandic followed by continuous fine-tuning on 100h faroes (ft→cft, e2). https://huggingface.co/davidilag/wav2vec2-xls-r-1b-E2-faroese-100h-30-epochs_20250124_v3.

Dávid í Lág. 2025d. Wav2vec2 xls-r 1b: Fine-tuning on 100h norwegian followed by continuous fine-tuning on 100h faroese (ft→cft, e4). https://huggingface.co/davidilag/wav2vec2-xls-r-1b-E4-faroese-100h-30-epochs_20250208_v4.

Dávid í Lág. 2025e. Wav2vec2 xls-r 1b: Fine-tuning on 100h swedish followed by continuous fine-tuning on 100h faroese (ft→cft, e5). https://huggingface.co/davidilag/wav2vec2-xls-r-1b-E5-faroese-100h-30-epochs_20250124.

Dávid í Lág. 2025f. Wav2vec2 xls-r 1b: Fine-tuning on 25h each of danish, icelandic, norwegian, and swedish followed by continuous fine-tuning on 100h faroese (ft→cft, e6). https://huggingface.co/davidilag/wav2vec2-xls-r-1b-E6-faroese-100h-30-epochs_20250209.

Dávid í Lág. 2025g. Wav2vec2 xls-r 300m: Continuous pre-training on 1000h faroese followed by fine-tuning on 100h faroese (cpt→ft, e1). https://huggingface.co/davidilag/wav2vec2-xls-r-300m-cpt-1000h_faroese-cp_best-faroese-100h-60-epochs_run8_2025-09-24.

Dávid í Lág. 2025h. Wav2vec2 xls-r 300m: Continuous pre-training on 200h each of faroese, danish, icelandic, norwegian, and swedish followed by fine-tuning on 100h faroese (cpt→ft, e6). https://huggingface.co/davidilag/wav2vec2-xls-r-300m-cpt-200h-FO-IS-NO-DK-SE-cp-best-faroese-100h-30-epochs_run9_2025-09-11.

Dávid í Lág. 2025i. Wav2vec2 xls-r 300m: Continuous pre-training on 500h faroese and 500h danish followed by fine-tuning on 100h faroese (cpt→ft, e3). https://huggingface.co/davidilag/wav2vec2-xls-r-300m-pt-500h-FO-500h-DK-cp-best-ft-faroese-100h-30-epochs_run9_2025-09-11.

Dávid í Lág. 2025j. Wav2vec2 xls-r 300m: Continuous pre-training on 500h faroese and 500h icelandic followed by fine-tuning on 100h faroese (cpt→ft, e2). https://huggingface.co/davidilag/wav2vec2-xls-r-300m-pt-500h-FO-500h-IS-cp-best-faroese-100h-30-epochs_run9_2025-09-10.

Dávid í Lág. 2025k. Wav2vec2 xls-r 300m: Continuous pre-training on 500h faroese and 500h norwegian followed by fine-tuning on 100h faroese (cpt→ft, e4). https://huggingface.co/davidilag/wav2vec2-xls-r-300m-pt-500h-FO-500h-NO-cp-best-faroese-100h-30-epochs_run9_2025-09-11.

Dávid í Lág. 2025l. Wav2vec2 xls-r 300m: Continuous pre-training on 500h faroese and 500h swedish followed by fine-tuning on 100h faroese (cpt→ft, e5). https://huggingface.co/davidilag/wav2vec2-xls-r-300m-pt-500h-FO-500h-SE-cp-best-faroese-100h-30-epochs_run9_2025-09-11.

10. Language Resource References

2023. *NST Norwegian ASR Database (16 kHz) – Reorganized*. Originally developed by Nordisk Språkteknologi; reorganized by the National Library of Norway (Språkbanken). Last updated 2023-12-19.

Rosana Ardila and Megan Branson and Kelly Davis and Michael Henretty and Michael Kohler and Josh Meyer and Reuben Morais and Lindsay Saunders and Francis M. Tyers and Gregor Weber. 2020. *Common Voice: A Massively-Multilingual Speech Corpus*.

Helgadóttir, Inga Rún and Kjaran, Róbert and Nikulásdóttir, Anna Björk and Gudnason, Jón. 2021. *Althingi Parliamentary Speech*. Reykjavík University. LDC Catalog No. LDC2021S01.

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*.

Dávid í Lág. 2025. *Scandinavian ASR Dataset: 100 Hours per Language (Danish, Icelandic, Norwegian, Swedish)*. Hugging Face. Public dataset for multilingual ASR research.

Nielsen, Dan Saattrup and Lehmann, Sif Bernstorff and Madsen, Simon Leminen and Pedersen, Anders Jess and van Zee, Anna Katrine and Blach, Torben. 2024. *CoRal: A Diverse Danish ASR Dataset Covering Dialects, Accents, Genders, and Age Groups*.

O'Brien, Luke and Gunnarsson, Thorsteinn Dadi and Magnúsdóttir, Eydís Huld and Gudnason, Jon. 2024. *Samrómur L2 24.10*. Reykjavík University.

Rekathati, Faton. 2023. *The KBLab Blog: RixVox: A Swedish Speech Corpus with 5500 Hours of Speech from Parliamentary Debates*.

Solberg, Per Erik and Ortiz, Pablo. 2022. *The Norwegian Parliamentary Speech Corpus*.

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.

í Lág, Dávid. 2025. *FPSC: Faroese Parliament Speech Corpus*. University of the Faroe Islands. Public dataset of Faroese parliamentary speech with metadata and weakly supervised transcripts, 1,600 hours.