Mapping Faroese in the Multilingual Representation Space: Insights for ASR Model Optimization

Dávid í Lág
University of the Faroe Islands
J. C. Svabosgøta 14,
100 Tórshavn
davidl@setur.fo

Barbara Scalvini
University of the Faroe Islands
J. C. Svabosgøta 14,
100 Tórshavn
barbaras@setur.fo

Jón Gunason
Reykjavik University
Menntavegur 1
101 Reykjavik
jb@ru.is

Abstract

ASR development for low-resource languages such as Faroese faces significant challenges due to the scarcity of large, diverse datasets. Although fine-tuning multilingual models using related languages is common practice, there is no standardized method for selecting these auxiliary languages, leading to a computationally expensive trial-and-error process. By analyzing the positioning of Faroese among other languages in wav2vec2's multilingual representation space, we find that Faroese's closest neighbors are influenced not only by linguistic similarity but also by historical, phonetic, and cultural factors. These findings open new avenues for auxiliary language selection to improve Faroese ASR and underscore the potential value of data-driven factors in ASR fine-tuning.

1 Introduction

Low-resource languages, such as Faroese, face unique challenges in ASR development, primarily due to the lack of sufficiently large and varied datasets. Recent advances in multilingual ASR models have provided a promising avenue for cross-linguistic transfer, leveraging similarities between languages to enhance the performance of those with limited resources. It is common practice to fine-tune multilingual models for a target language by incorporating similar, closely related languages (Juan et al., 2014; Juan, 2015; Ivan Froiz-Mıguez, 2023). However, currently there is no standardized procedure for selecting these languages. ASR researchers often train multiple models with different language combinations to find the best set to enhance target language performance, a trial-and-error approach that is computationally costly as models grow larger. This underscores the need for more efficient methods. In this study, we focus on Faroese, a low-resource Insular Scandinavian language. We explore its representation in Meta's wav2vec2 XLS-R 53 model (Alexis Conneau, 2020), and seek out its neighbors in this space, with the aim of extracting new insight for selection of auxiliary languages. Our approach analyzes how languages are encoded within the model's multilingual representation space by measuring the distance between Faroese and 102 languages from the Google Fleurs dataset (Alexis Conneau, 2022) at each model layer. Since Faroese is absent from Google Fleurs, we incorporated recordings from the Ravnursson data set (Hernández Mena and Simonsen, 2022), currently the only ASR-suitable Faroese dataset, to better understand how the model perceives Faroese in relation to other languages and to improve multilingual fine-tuning strategies.

2 Background and related work

2.1 Advances in Transformer and Self-Supervised Models for ASR

In 2019, the wav2vec model was introduced as a self-supervised model that learns speech representations without labeled data and can be fine-tuned for ASR, reducing the need for extensive labeled datasets (A. Baevski and Auli, 2020). While initially trained only on English, later versions support multiple languages (Alexis Conneau, 2020). The architecture of the wav2vec 2.0 model enables cross-lingual transfer in ASR through multilingual quantized speech representations, allowing latent speech units to capture key features of speech (Alexei Baevski, 2020). Transfer learning with related languages has been shown to improve ASR for low-resource languages by leveraging high-dimensional embeddings from the wav2vec2.0 XLSR-53 model (Akbayan Bekarystankyzy, 2024; J. Cho and Hori, 2018; Vishwa Gupta, 2022). Re-

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 354–358
March 3-4, 2025

search demonstrates the model's ability to capture language similarities by clustering embeddings using K-Means (Alexis Conneau, 2020).

2.2 ASR for Faroese

The effort towards digitalization of Faroese speech has led to the creation of a Basic Language Resource Kit for Faroese (A. Simonsen and Henrichsen, 2022) in the context of the Ravnur project.[1] This project involved the collection of both text corpora and audio recordings finalized in the creation of ASR systems. The Ravnur audio data set contains 100 hours of training data, which is a balanced collection of high-quality recordings, including different dialects and speakers of different ages. The availability of such data has allowed researchers to test strategies to produce ASR models for Faroese. One such strategy was the fine-tuning of multilingual models such as wav2vec2, which led to the creation of the very first ASR model specifically targeting Faroese (Hernandez Mena, 2022).

3 Method

3.1 Dataset

To assess the relationship between Faroese and other languages, we used Meta's wav2vec2 XLS-R 53 Large model[2] with 25 layers to generate hidden representations for all of the 102 Google Fleurs[3] (Alexis Conneau, 2022) languages in addition to Faroese. The model is trained on 56k hours of speech data for 53 languages. Of the Scandinavian languages, only Swedish is included in the model. We performed inference with the model using the same number of sentences per language in the Google Fleurs dataset for the 102 languages. Faroese is not in Google Fleurs, and therefore we instead take 900 random sentences from the Ranvursson ASR corpus[4].

3.2 Distance calculation

We calculate the distance between Faroese and 102 other languages in the hidden representation space of wav2vec 2.0, analyzing across different layers. The pipeline for the distance calculation can be summarized as follows. First, we obtain a sentence-level representation by applying average pooling to all hidden representations across the sentence. Then, we compute the overall representation by averaging the sentence-level representations for all sentences for each language l and layer j,

$$\mu_{l,j} = \frac{1}{N} \sum_{i=1}^{N} R_{l,i,j}, \tag{1}$$

where $R_{l,i,j}$ is the representation vector for sentence s_{li} at layer j. $S_l = s_{l1}, s_{l2}, \ldots, s_{lN}$ is a set of $N = 900$ sentences for language $l \in L$ where L is a set of languages with $|L| = 103$. The layer index is $j = 0, 1, \ldots, 24$.

3.3 Clustering and visualization

K-means clustering was used on the computed representations after performing dimensionality reduction using Principal Component Analysis (PCA) (Jolliffe, 2002), t-distributed stochastic neighbor embedding (t-SNE) (T. Tony Cai, 2021) and Uniform Manifold Approximation and Projection (UMAP) (Leland McInnes, 2018). Each layer in the wav2vec2 XLS-R 53 model contributes to the model's overall functionality. Ankita Pasad (2021) explored which type of speech information is predominantly encoded in each of the 25 layers of the wav2vec2 model, in terms of local acoustic features, phone identity, word identity, and word meaning. We take inspiration from their results and identify three main layer groupings:

- Layers 1 to 11: The first few layer representations (0-5) are dominated by local acoustic features, which gradually decrease, leaving gradually room for language-specific features such as phone and word identity.
- Layers 12 to 19: In these layers, word identity and word meaning dominate the representations, capturing more abstract linguistic features essential for understanding syntax and semantics. There is a sharp decrease in phone identity representation around layer 15, followed by a sharp increase.
- Layers 20 to 24: We observe an overall decrease in all linguistic properties, with phone identity, however, remaining more prominent than the other characteristics.

[1] https://mtd.setur.fo/en/resource/ravnur-blark-1-0/

[2] https://huggingface.co/facebook/wav2vec2-large-xlsr-53

[3] https://huggingface.co/datasets/google/fleurs

[4] https://huggingface.co/datasets/carlosdanielhernandezmena/ravnursson_asr

We use this information for interpretations of the results and layer selection during qualitative clustering analysis. Specifically, we will focus on layers 18 - 20, as we expect word identity and phone identity information to be at their highest in these layers.

3.4 Experiments

The key steps involved in our methodology are outlined as follows:

1. **Data selection:** Since Icelandic had the fewest sentences in the Google Fleurs dataset, with 924 sentences, we set the number of sentences per language for the analysis at 900.

2. **Hidden representation extraction:** For each language, we ran inference with the wav2vec2 XLS-R 53 model on the selected 900 sentences, extracting the hidden representation for each of the 25 hidden layers as described in Sec 3.3. We processed the representations as follows:

    - Calculating the mean of all layer-wise 25 hidden representations per language
    - Grouping the layers into intervals of five: 0-4, 5-9, 10-14, 15-19, 20-24, and computing the mean interval representation for each language.

3. **Distance between languages:** To explore the relationships between Faroese and the other languages, we calculate the Euclidean distance in the original representation vector space.

4. **Clustering:** We apply K-Means after reducing dimensions down to 2 using PCA, t-SNE, and UMAP. This choice was made in order to facilitate visualization and qualitative analysis.

4 Results and Discussion

4.1 Quantitative analysis: top nearest neighbors in the representation space for Faroese

For each layer interval, we calculated the Euclidean distance between Faroese and the 102 languages in the Google Fleurs dataset. Table 1 presents the top eight nearest neighbors to Faroese in descending order for each layer interval. Interesting patterns emerge from these results. The top nearest neighbor across all layer intervals is either Welsh or Irish, with Welsh being the closest when all layers (0–24) are combined. Welsh and Irish belong to the Celtic language family, in contrast to Faroese, which is a Scandinavian language. However, Faroese phonetics is known to have been significantly influenced by contact with Scottish Gaelic-speaking communities from the neighboring British Isles. German ranks as the second closest neighbor in the early layers (0–9), while Scandinavian languages emerge as neighbors in the later layers: Swedish in layers 10–14, and Norwegian in layers 20–24 and overall. Beyond this, the composition of nearest neighbors does not reveal any clear pattern in terms of linguistic families.

4.2 Qualitative analysis: dimensionality reduction and clustering

The internal representation space of multilingual models is highly multidimensional and often challenging to interpret. To clarify the results of our quantitative analysis and provide a visual interpretation of the distances in this space, we performed dimensionality reduction on the combined representation space of layers 18–20. In these layers, we anticipate clustering among languages from the same linguistic families due to shared phonetic, syntactic, or acoustic characteristics. If a language clusters separately from its family, it may indicate unique linguistic traits. Examining outliers and mixed clusters could also uncover cross-family influences or reveal features such as geographic convergence. Figure 1 shows clusters of languages in the same language family for six different regions. Clustering was performed using K-Means following dimensionality reduction to two dimensions. Of the three-dimensionality reduction techniques tested, t-SNE most closely aligned with results from the original high-dimensional space, as shown in Table 2. In this analysis, Irish appears as the closest neighbor to Faroese, with Swedish positioned farther within the neighborhood (see Figure 1). Overall, we observe a representation of Germanic/Scandinavian languages in the clusters (English, German, Luxembourgish, Swedish), along with non-Indo-European languages that are part of the Nordic cultural sphere, such as Finnish.

Layers	0-4	5-9	10-14	15-19	20-24	0-24
1	Irish (10.8)	Irish (13.4)	Irish (13.4)	Irish (16.2)	Welsh (31.7)	Welsh (14.8)
2	German (11.3)	German (15.4)	Estonian (15.4)	Croatian (17.0)	Turkish (34.7)	Turkish (17.5)
3	Romanian (11.6)	Estonian (16.0)	Croatian (15.8)	Estonian (17.4)	Punjabi (47.4)	Punjabi (22.6)
4	Estonian (11.8)	Croatian (16.2)	Lithuanian (15.9)	Lithuanian (17.5)	Slovak (104.0)	Slovak (25.2)
5	Simplified Chinese (11.8)	Romanian (16.2)	Welsh (16.1)	Polish (17.7)	Georgian (110.1)	Georgian (25.8)
6	Catalan (12.0)	English (16.2)	Romanian (16.1)	Georgian (17.9)	Amharic (112.7)	Amharic (27.4)
7	Korean (12.1)	Welsh (16.4)	Polish (16.5)	Romanian (18.0)	Norwegian (126.4)	Norwegian (29.8)
8	Armenian (12.3)	Lithuanian (16.4)	Swedish (16.6)	Slovenian (18.0)	Vietnamese (145.8)	Armenian (32.5)

Table 1: *Closest languages to Faroese measured in Euclidean distance in the original representation vector space*

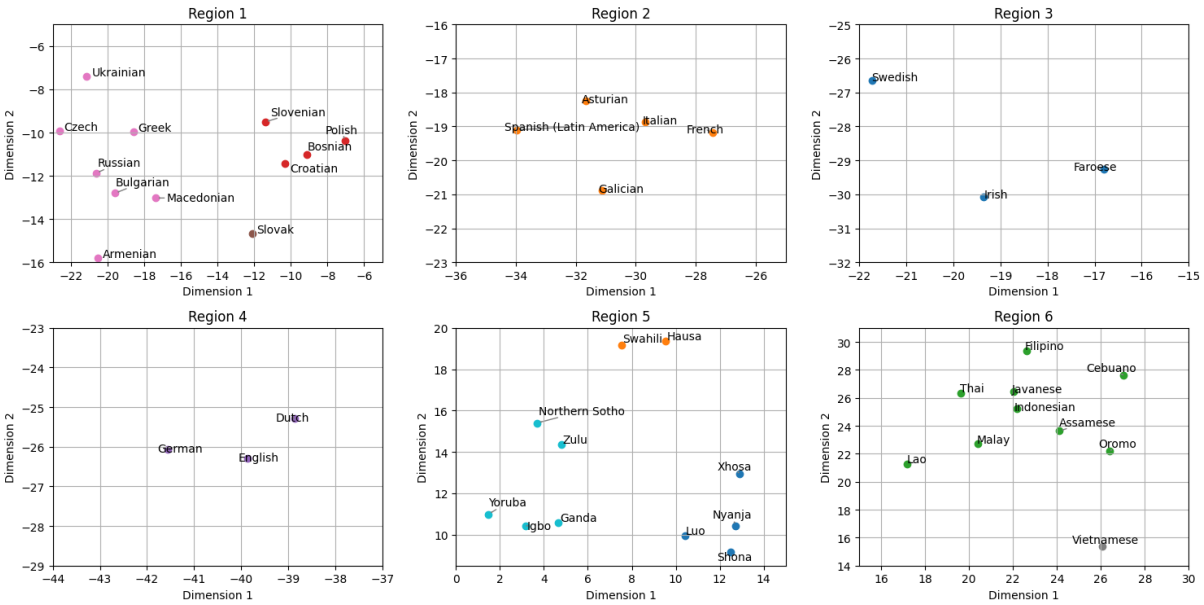

Figure 1: *Clusters of closely related languages for layers 18-20 with t-SNE and K-Means with 18 clusters*

PCA	t-SNE	UMAP
Romanian (1.72)	Irish (2.68)	Croatian (0.42)
French (3.29)	Maori (4.03)	Catalan (0.47)
English (5.51)	Swedish (5.58)	Romanian (0.55)
German (5.64)	Finnish (6.00)	Maori (0.77)
Luxembourgish (9.83)	Latvian (8.14)	Georgian (0.79)

Table 2: *Languages in the same cluster as Faroese in layers 18-20 using K-Means with 18 clusters after dimensional reduction with PCA, t-SNE, and UMAP*

5 Conclusion

In conclusion, the representation spaces in wav2vec2 indicate that languages tend to clusters, as evidenced through nearest-neighbor analysis, clustering, and dimensionality reduction techniques. This analysis places Faroese in proximity to Gaelic languages, alongside Germanic and Nordic languages. The prominence of Gaelic languages as close neighbors suggests that limiting comparisons to only the closest family members may overlook valuable insights, possibly related to historical phonetic and linguistic influences. Such consideration will be further investigated in future work.

6 Limitations

This exploration of the representation of Faroese is based on a single model and may therefore vary with other models, as language representations are influenced by the specific language distribution

within the training data. Additionally, we only evaluated language proximity using one dataset, FLEURS, which may have limited speaker representation. The metric used, Euclidean distance, is just one approach for vector comparison and has its limitations. For instance, it is susceptible to the curse of dimensionality and may not be optimal in highly multidimensional spaces. Alternative metrics, such as cosine similarity, could yield slightly different results. Despite these limitations, our analysis provides a foundation for a more comprehensive characterization of language similarity within model representation spaces, with potential applications in language selection for low-resource ASR training.

References

A. Mohamed A. Baevski, H. Zhou and M. Auli. 2020. http://arxiv.org/abs/2006.11477 wav2vec 2.0: A framework for self-supervised learning of speech representations.

I. N. Debess A. Simonsen, S. S. Lamhauge and P. J. Henrichsen. 2022. https://aclanthology.org/2022.lrec-1.495/ Creating a basic language resource kit for faroese.

Mateus Mendes Anar Fazylzhanova Muhammad Assam Akbayan Bekarystankyzy, Orken Mamyrbayev. 2024. https://www.nature.com/articles/s41598-024-64848-1 Multilingual end-to-end asr for low-resource turkic languages with common alphabets.

Michael Auli Alexei Baevski, Alexis Conneau. 2020. https://ai.meta.com/blog/wav2vec-20-learning-the-structure-of-speech-from-raw-audio/ Wav2vec 2.0: Learning the structure of speech from raw audio.

Ronan Collobert Abdelrahman Mohamed Michael Auli Alexis Conneau, Alexei Baevski. 2020. https://arxiv.org/abs/2006.13979 Unsupervised cross-lingual representation learning for speech recognition.

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

Karen Livescu Ankita Pasad, Ju-Chieh Chou. 2021. https://arxiv.org/abs/2107.04734 Layer-wise analysis of a self-supervised speech representation model.

Carlos Daniel Hernandez Mena. 2022. https://huggingface.co/carlosdanielhernandezmena/wav2vec2-large-xlsr-53-faroese-100h Acoustic model in faroese: wav2vec2-large-xlsr-53-faroese-100h.

Carlos Daniel Hernández Mena and Annika Simonsen. 2022. http://hdl.handle.net/20.500.12537/276 Ravnursson faroese speech and transcripts.

Paula Fraga-Lamas Diego Fustes Carlos Dafonte Javier Pereira Tiago M. Fernandez-Carames Ivan Froiz-Mıguez, Oscar Blanco-Novoa. 2023. https://doi.org/10.29007/1ppr Design and evaluation of a cross-lingual ml-based automatic speech recognition system fine-tuned for the galician language. *Kalpa publications in computing*.

R. Li M. Wiesner S. H. Mallidi N. Yalta M.Karafiat S. Watanabe J. Cho, M. K. Baskar and T. Hori. 2018. https://arxiv.org/abs/1810.03459 Multilingual sequence-tosequence speech recognition: architecture, transfer learning, and language modeling. *2018 IEEE Spoken Language Technology Workshop (SLT)*, arXiv:1810.03459.

Ian T Jolliffe. 2002. *Principal component analysis for special types of data*. Springer.

Sarah Flora Samson Juan. 2015. https://api.semanticscholar.org/CorpusID:33165732 Exploiting resources from closely-related languages for automatic speech recognition in low-resource languages from malaysia.

Sarah Flora Samson Juan, Laurent Besacier, Benjamin Lecouteux, and Tien Ping Tan. 2014. https://api.semanticscholar.org/CorpusID:8620301 Using closely-related language to build an asr for a very under-resourced language: Iban. *2014 17th Oriental Chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques (COCOSDA)*, pages 1–5.

James Melville Leland McInnes, John Healy. 2018. https://arxiv.org/abs/1802.03426 Umap: Uniform manifold approximation and projection for dimension reduction.

Rong Ma T. Tony Cai. 2021. https://arxiv.org/abs/2105.07536 Theoretical foundations of t-sne for visualizing high-dimensional clustered data.

Gilles Boulianne Vishwa Gupta. 2022. https://aclanthology.org/2022.lrec-1.689.pdf Progress in multilingual speech recognition for low resource languages kurmanji kurdish, cree and inuktut.