Interspeech 2025
17-21 August 2025, Rotterdam, The Netherlands

Is your model big enough? Training and interpreting large-scale monolingual speech foundation models

Yaroslav Getman[1], *Tamás Grósz*[1], *Tommi Lehtonen*[2], *Mikko Kurimo*[1]

[1]Department of lnformation and Communications Engineering, Aalto University, Finland
[2]Finnish National Audiovisual Institute (KAVI), Finland
firstname.lastname@aalto.fi, tommi.lehtonen@kavi.fi

Abstract

Self-supervised learning has been widely used in developing speech foundation models. Most languages, however, are only represented in multilingual foundations. We introduce monolingual self-supervised foundation models pre-trained on more than 150,000 hours of Finnish speech and propose a new interpretation technique to understand their capabilities. To our knowledge, this is the largest monolingual data used for self-supervised non-English speech representation learning. Our models demonstrate superior downstream low-resource ASR performance and improved generalization compared to prior work, with absolute WER reductions of up to 14%. Moreover, our proposed interpretation technique, Layer Utilization Rate (LUR), enables us to assess the percentage of neurons in each layer highly contributing towards the output. Empirical results show that the proposed LUR metric can be used to indicate the potential of the fine-tuned model's size and architecture to generalize to unseen domains.

Index Terms: wav2vec2, ASR, pre-training, interpretation

1. Introduction

Self-supervised speech representation learning (SSL) has become standard in many fields, including Automatic Speech Recognition (ASR). The so-called foundation models efficiently leverage large amounts of unlabeled speech and demonstrate high performance both as feature extractors and downstream models. Unfortunately, most available models are pre-trained using English-only or multilingual data, giving limited options for other languages such as Finnish. While multilingual models fine-tuned with Finnish data have shown good results [1, 2], recent studies report that properly trained monolingual models outperform multilingual ones [3]. This raises the question of how much performance can be further enhanced with monolingual foundations. Another issue is that when ASR models are used on out-of-domain data, as often happens in practice, they perform poorly [4–6]. Addressing these challenges by developing a foundation model from scratch requires access to both vast amounts of diverse unlabeled speech data and substantial computational resources.

The most notable among the monolingual foundations is the series of VoxPopuli models [7] pre-trained on European Parliament plenary session recordings, providing speech foundation models for 23 languages. However, only smaller models are monolingual and the pre-training data lacks diversity: it consists of formal speech from a relatively small number of speakers. [8] presents a series of foundation models pre-trained on 1K to 14K hours of French. In [9], a 95M-parameter model was pre-trained on 80K hours of Czech, while 11K hours were utilized to train a 317M-parameter Swedish model in [10].

Inspired by these works, we aim to fill this gap for Finnish by introducing three new monolingual speech foundation models with up to 1B parameters pre-trained on more than 150,000 hours of Finnish speech. To the best of our knowledge, it is the largest corpus used to develop open-source monolingual non-English SSL speech models.

While ASR fine-tuned foundation models are usually compared using metrics such as word error rate (WER), deeper analysis is needed to understand how they actually work and why some perform better than others. Previous works mostly focused on probing for specific knowledge [11–13], evaluating various hidden layers [14–17] or the codebooks learned during the self-supervised phase [18]. In contrast, we adopt a fundamentally different approach based on explainable AI (XAI) methods [19], specifically integrated gradients [20]. This paradigm shift emphasizes model interpretability and explanation at a more fine-grained (neuron) level rather than relying on probing and layer-wise analysis, which treat hidden layers as atomic units. In this paper, we propose a novel post hoc interpretation metric called Layer Utilization Rate (LUR), which measures the percentage of actively contributing neurons in each hidden layer. To separate non-contributing and vital neurons, we utilize a gradient-based attribution technique and propose a simple thresholding solution for categorizing hidden neurons in the model.

To summarize, the main contributions of this work are two-fold. First, we present a series of monolingual Finnish self-supervised foundation models pre-trained on 157,800 hours of speech from a diverse set of databases as well as their downstream ASR versions fine-tuned on nearly 4,600 hours of Finnish. Second, we introduce a novel metric for analyzing fine-tuned models and discovering their internal capabilities. We share the new models and the scripts with the community via GitHub[1] and HuggingFace Hub[2].

2. Data

Table 1 summarizes the unlabeled pre-training data we used. Nearly 90% of it consisted of recordings of TV and radio broadcasts from the Radio and Television Archive (RTVA), provided by the Finnish National Audiovisual Institute (KAVI). In total, 16 radio and 11 TV channels have been archived on a 24/7 basis since 2009. The best source of clearly spoken Finnish content is the talk radio station YlePuhe. For more casual and variable content, we got fixed time slots featuring predominantly Finnish-speaking programs from three TV channels.

[1]https://github.com/aalto-speech/large-scale-monolingual-speech-foundation-models

[2]https://huggingface.co/collections/GetmanY1/wav2vec2-fi-150k-66c9d75d18579088974ea37f

10.21437/Interspeech.2025-46

Subset	Domain	Size		
		Hours		Samples
		Raw	After VAD	
YlePuhe	Radio	107.8K	95.5K	29.8M
MTV3	TV	25.5K	16.7K	7.1M
YleTV1	TV	20.8K	13.9K	5.9M
AlfaTV	TV	20.4K	11.9K	4.3M
VoxPopuli	Parl.	N/A	14.3K	1.8M
FinParl	Parl.	3.1K	2.7K	1.8M
LP	Colloq.	3.2K	2.7K	0.5M
Total		195.1K	157.8K	51.2M

Table 1: *Data used for pre-training.*

The remaining data consisted of publicly available Finnish speech datasets. First, we included the Finnish subset of VoxPopuli [7] (VP) and the Finnish Parliament dataset (FinParl) [2]. The latter is fully transcribed and was applied also for ASR fine-tuning. Lastly, we used the collection of colloquial Finnish speech namely Lahjoita puhetta (Donate Speech, LP) [1]. The corpus is partially transcribed, and its annotated part (1.6K hours) was used also for fine-tuning.

To split the data into shorter utterances and reduce the non-speech events, such as music, noise, and silence, we pre-processed all the data except VoxPopuli (VAD-based segmentation performed by [7]) with a neural voice activity detector (VAD) [21]. Those resulting speech segments exceeding 60 seconds were further manually divided into shorter samples. The pre-processing reduced the size of the pre-training data by 19%, with most non-speech detected in the TV data.

3. Pre-training Framework

In this work, we opted for wav2vec 2.0 [22], an SSL framework widely used for mono- and multilingual speech representation learning in languages other than English. This choice not only enabled fairer comparison with existing foundations but also provided deeper insights into optimal hyperparameter selection.

Following [23], we opted for 3 different wav2vec 2.0 architecture sizes: *Base* with 12 Transformer layers and 95M parameters, *Large* consisting of 24 Transformer layers and 317M parameters, and *X-Large* composed of 48 Transformer layers and 965M parameters.

4. Layer Utilization Rate

In order to better understand our models, we opted to use tools of explainable artificial intelligence (XAI) [19], specifically layer and neuron attribution estimations. Our primary goal was to estimate the percentage of contributing neurons, thus confirming whether the fine-tuned model was optimally used in this task. For our analysis, we selected the Integrated Gradients (IG) technique [20], which estimates the individual neurons' contributions towards the final output using a gradient-based technique. In our recent work, we demonstrated that IG is a reliable tool for selecting task-relevant subspaces of various pre-trained foundation models [24], and here, we take it a step further by extending the analysis to the ASR task and all hidden layers.

To calculate neuron attributions using IG, we start with an input x and a corresponding baseline input x'. In our experiments, the baseline is a completely silent audio, represented as a vector of zeros. Next, a smooth path $\gamma : [0, 1] \rightarrow R^N$ is created between the baseline and the actual input ($\gamma(0) = x'$ and $\gamma(1) = x$). To get the neural attributions of a certain layer, we can split the model, $\Theta(x) = f(g(x))$, where $g(x)$ is the output of the layer that we wish to investigate and f is the remaining upper part of the network. Using this notation, the layer attributions of model Θ are defined as

$$IG(x, f, g, \gamma)_i = \int_{\alpha=0}^{1} \frac{\partial f(g(\gamma(\alpha)))}{\partial g(\gamma(\alpha))_i} \frac{\partial g(\gamma(\alpha))_i}{\partial \alpha} \, d\alpha, \quad (1)$$

where $\frac{\partial f(g(x))}{\partial g(x)_i}$ is the gradient of the model's output with respect to the i^{th} neuron of the layer. For more technical details about IG, we refer the interested reader to [20] and [25].

After estimating the neuron attributions, we can separate the valuable neurons from those with low values. Specifically, we estimate the neuron attributions for each output symbol separately. Based on our preliminary observation that the attributions approximately follow a normal distribution, we define all neurons with attributions within one standard deviation from the mean as non-contributing. Finally, we look for neurons in the non-contributing set of all output symbols. These units are classified as irrelevant for a given task, and the layer utilization rate is defined as $LUR = 1 - \frac{\#irrelevant\ neurons}{\#neurons}$

5. ASR Experiments

In this work, we trained the foundation models with the Fairseq toolkit [26] on the LUMI supercomputer [27]. Each model was pre-trained on 256 AMD MI250x modules (512 GPUs). The *Base* model was trained with the batch size of 25 hours (3 min./GPU) for 125K steps (16 epochs). For the *Large* one, the total batch size was 16 hours (2 min./GPU) and it was trained for 167K steps (14 epochs). Finally, the *X-Large* model was trained with the batch size of 8.5 hours (1 min./GPU) for 390K steps (19 epochs). The full list of hyperparameters can be accessed from GitHub[3]. Training took around 1, 5, and 10 days for the *Base*, *Large*, and *X-Large* models, respectively. 2 hours of each data subset were held out for validation. We refer to our pre-trained models as KAVI.

Next, we compared the ASR performance of our models to publicly available foundations that included Finnish speech during pre-training, namely the monolingual *Base* and *Large* ones pre-trained on the Lahjoita puhetta data (LP) [18] and Finnish *Base*-sized VoxPopuli (VP) model [7] as well as multilingual Uralic VoxPopuli (*Large*) [7] and XLS-R (*Large* and *X-Large*) [23]. For diverse comparison in terms of data scenarios, we started with a low-resource ASR experiment by fine-tuning them either on Common Voice 16.1 (2.7 hours) [28] or FLEURS (8.8 hours) [29] and evaluating on the corresponding development and test sets. The models were fine-tuned using HuggingFace Transformers [30] on 8 GPUs for 80 epochs.

High-resource ASR fine-tuning was performed with the Finnish Parliament dataset and the labeled part of the Lahjoita puhetta (LP) corpus. LP contained long utterances which were split into shorter samples with silence removal, reducing its total duration to 1,500 hours. As a result, the overall size of the labeled training data was nearly 4,600 hours. Fine-tuning was done with Fairseq on 128 GPUs with a total batch size of 5.3 hours for the *Base* and *Large* models and 256 GPUs with a batch size of 4.6 hours for the *X-Large* models for 15 epochs. For a

[3]https://github.com/aalto-speech/large-scale-monolingual-speech-foundation-models

System	Training and Evaluation Data			
	Common Voice		FLEURS	
	Dev	Test	Dev	Test
Base Models				
VP-Fin	28.5	$26.6^{(29.9}_{23.6)}$	18.3	$19.4^{(20.4}_{18.4)}$
LP	29.9	$27.4^{(30.4}_{24.9)}$	22.7	$22.7^{(23.7}_{21.7)}$
KAVI	16.1	$14.2^{(16.4}_{12.1)}$	12.3	$13.6^{(14.5}_{12.8)}$
Large Models				
VP-Uralic	17.9	$16.7^{(19.1}_{14.6)}$	16.0	$16.7^{(17.6}_{15.8)}$
XLS-R	30.5	$26.8^{(29.8}_{23.8)}$	22.2	$22.5^{(23.4}_{21.5)}$
LP	14.3	$13.1^{(14.9}_{11.3)}$	19.4	$19.9^{(20.9}_{19.0)}$
KAVI	8.5	$7.7^{(9.0}_{6.6)}$	13.1	$13.3^{(14.1}_{12.5)}$
X-Large Models				
XLS-R	20.1	$18.4^{(20.7}_{16.2)}$	13.6	$13.6^{(14.4}_{12.8)}$
KAVI	6.1	$6.3^{(7.4}_{5.2)}$	7.9	$8.6^{(9.2}_{7.9)}$

Table 2: *Low-resource ASR results in terms of WER [%] with 95% CIs after fine-tuning on Common Voice (2.7h) or FLEURS (8.8h), evaluated on the corresponding dev and test sets.*

comprehensive assessment of model generalizability, we evaluated the ASR performance in terms of word error rate (WER) on both in-domain and out-of-domain data, reporting the results along with the 95% confidence intervals (CIs) computed via bootstrapping [31].

Table 2 demonstrates that our new KAVI models consistently provided the best results in the low-resource ASR scenarios. The relative and absolute improvements over corresponding second-best models of the same size varied between 18%-69% and 2.9%-13.9% respectively, and a reduction in the performance gap can be observed with the increase of the labeled training data size. The 317M-parameter XLS-R provided particularly high error rates on Common Voice, indicating that its universal acoustic representations pre-trained on 128 languages failed to adapt to the target language with limited data.

High-resource ASR results summarized in Table 3 follow the trend of previous experiments and confirm the superiority and improved generalization of the new models, especially on out-of-domain data, with 12%-61% relative (1.0%-11.5% absolute) WER reduction (WERR) over second-best results. The models pre-trained solely on colloquial Finnish (LP) consistently ranked second on the corresponding evaluation data. Similarly, those including parliamentary speech for pre-training (VP and XLS-R) achieved higher performance on FinParl.

6. Discussion and Analysis

The diverse KAVI data proved to be the most beneficial, demonstrating once more that larger quantities of in-language pre-training data lead to more robust foundation models. Interestingly, in the case of sizable supervised data, we observe a considerable gap between the *Base* and *Large* models' performance. However, further increasing the number of parameters results in only minor improvements, which require further investigation.

Table 4 presents the inter-annotator disagreements between human transcribers and our *Large* and *X-Large* models, evaluated using WER on a separate 1-hour LP test set. If considering the error rates between the two most consistent human experts as an upper limit of human performance on conversational Finnish speech, the *Large* model is slightly behind, while the *X-Large* one matches human performance, marginally outperforming the top two transcribers, though the difference is not statistically significant.

While our ASR results demonstrated the superior performance of monolingual models pre-trained on large amounts of Finnish data, the reasons for their significantly improved performance are not yet fully understood. Using our proposed LUR method, we first took a closer look at the *Base*-sized models trained in the high-resource setting, specifically their attributions on the FLEURS test set, see Figure 1 for the LUR values of each layer. The first interesting observation is a common trend of higher layers showing higher utilization. This can be explained by the observation that lower parts of the Transformer are responsible for acoustic feature extraction, while higher layers combine the acoustic information with linguistic knowledge as reported in [15]. Thus, they require more actively contributing neurons, generally after the 6. layer (out of the 12 layers of the *Base* model). While looking at the overall average utilization rates of the three models we also observed that Base-VP had considerably fewer contributing units (LUR: 45.66%) than the others, especially in the lower parts, suggesting its acoustic features are more specialized for parliamentary speeches and less suited for this domain, which explains its inferior WER performance compared to the best model. In contrast, Base-LP has demonstrated considerably high overall LUR (85.75%), which implies it has overfitted to the single domain, namely colloquial Finnish in the LP corpus, and failed to generalize well for out-of-domain data like FLEURS resulting in the worst WER out of the three models. Lastly, the new Base-KAVI had an average LUR of 67.00%, yielding the best recognition rates. The acoustic parts LUR fluctuated between 50-60%, suggesting that it has managed to maintain a diverse acoustic feature extraction capability after fine-tuning that generalized well for new domains. Still, the last two layers had very high utilization, indicating that the model could benefit from additional units and layers.

Next, we compared different model sizes to analyze the suitability of architectures. The high LUR of *Base* models' last layers suggests that the vast majority of the units focused on recognizing characters while ignoring other aspects of the acoustic signal. Similarly, the *Large* models had lower LUR values for the first 6 layers, then a gradually increasing utilization rate until the last one. Based on this, we hypothesize that around 6 layers are needed for the acoustic information extraction, after which, the *Large* model could combine the linguistic knowledge at a slower pace, leading to considerably better WER (see Table 3). We should note that the last layers' LUR peaked around 90%, suggesting that the model had the appropriate capability and did not excessively overspecialize the final part of the Transformer for the ASR task. Looking at the *X-Large* models, we can see again that the acoustic part had lower LUR, while the linguistic parts' LUR gradually increased, peaking around 80-85%. Interestingly, after the 32. layer, a sharp drop in LUR can be observed, which might be a sign that the model was able to simplify the task to such a degree that only a portion of its units were needed for the ASR task. This indicates that the *X-Large* model architecture might be too big for the task and suffer from some overfitting, which explains why the gap between the *Large* and *X-Large* models is considerably smaller than between the *Base* and *Large* ones.

It is important to note that in our current experiments, we employed an adaptive threshold based on per-symbol attribution distributions to differentiate between contributing and non-contributing neurons. Although this approach accounts for vari-

System	In-domain					Out-of-domain			
	LP		FinParl			Common Voice		FLEURS	
	Dev	Test	Dev16	Test16	Test20	Dev	Test	Dev	Test
Base Models									
VP-Fin	29.7	$31.6^{(35.8}_{27.8)}$	17.0	$12.7^{(14.9}_{10.4)}$	$10.6^{(13.1}_{8.6)}$	19.1	$16.8^{(18.9}_{14.2)}$	16.1	$17.6^{(18.7}_{16.5)}$
LP	23.1	$25.5^{(29.1}_{22.6)}$	20.4	$15.6^{(17.8}_{13.2)}$	$13.0^{(15.6}_{10.6)}$	22.1	$18.7^{(20.9}_{16.3)}$	19.5	$20.9^{(22.0}_{19.8)}$
KAVI	20.0	$22.3^{(25.9}_{19.4)}$	15.0	$10.9^{(12.9}_{8.8)}$	$8.4^{(10.6}_{6.7)}$	12.5	$10.6^{(12.3}_{8.9)}$	12.0	$13.3^{(14.2}_{12.4)}$
Large Models									
VP-Uralic	20.8	$22.6^{(26.1}_{19.8)}$	13.6	$10.5^{(12.1}_{8.8)}$	$7.8^{(10.1}_{6.3)}$	12.8	$10.9^{(12.6}_{9.1)}$	12.8	$14.1^{(15.1}_{13.1)}$
LP	18.4	$20.2^{(23.3}_{17.6)}$	15.1	$11.2^{(13.1}_{9.2)}$	$9.2^{(11.4}_{7.3)}$	15.3	$13.4^{(15.3}_{11.4)}$	14.7	$16.5^{(17.5}_{15.5)}$
XLS-R	21.0	$22.9^{(26.2}_{20.2)}$	14.6	$10.9^{(12.9}_{8.9)}$	$8.8^{(11.1}_{6.9)}$	13.7	$12.1^{(13.9}_{10.2)}$	13.5	$15.2^{(16.3}_{14.3)}$
KAVI	15.3	$16.9^{(19.9}_{14.5)}$	11.3	$8.3^{(9.9}_{6.7)}$	$6.9^{(8.9}_{5.3)}$	7.2	$5.9^{(7.0}_{4.9)}$	9.2	$10.7^{(11.6}_{9.8)}$
X-Large Models									
XLS-R	26.3	$27.9^{(31.1}_{25.0)}$	13.6	$10.4^{(12.0}_{8.7)}$	$8.0^{(10.2}_{6.3)}$	16.0	$13.8^{(15.6}_{11.9)}$	14.9	$16.1^{(17.1}_{15.0)}$
KAVI	15.0	$16.4^{(19.2}_{14.1)}$	10.9	$7.8^{(9.3}_{6.3)}$	$6.4^{(8.4}_{4.9)}$	6.7	$5.4^{(6.5}_{4.4)}$	8.7	$10.0^{(10.8}_{9.2)}$

Table 3: *ASR results in terms of WER [%] with wav2vec 2.0 models fine-tuned on LP and FinParl, with 95% CIs*

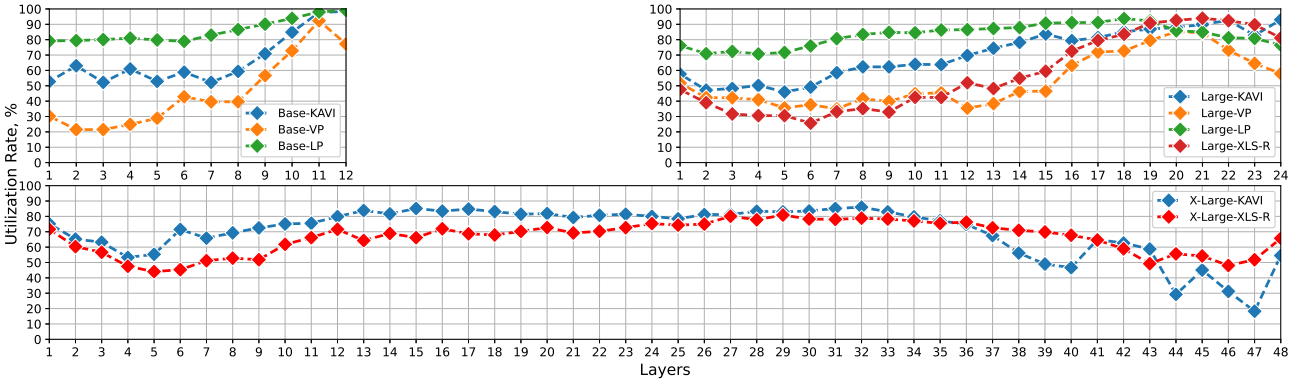

Figure 1: *LUR curves of the high-resource scenario models (see Table 3).*

	T2	T3	T4	KAVI-L	KAVI-XL
T1	$19.5^{(24.3}_{14.1)}$	$19.8^{(24.4}_{14.8)}$	$20.5^{(25.4}_{15.1)}$	$21.1^{(26.0}_{15.1)}$	$20.8^{(25.8}_{14.9)}$
T2	-	$\mathbf{13.6}^{(17.6}_{11.0)}$	$15.6^{(20.2}_{13.0)}$	$14.4^{(18.4}_{11.9)}$	$\mathbf{13.5}^{(18.1}_{10.9)}$
T3	-	-	$16.0^{(19.9}_{13.5)}$	$16.6^{(22.1}_{13.2)}$	$15.6^{(21.3}_{12.1)}$

Table 4: *Inter-annotator disagreements between human transcribers (T) and our Large and X-Large models in terms of WER [%] with 95% CIs on the 1-hour LP test set.*

ability across output symbols, future work could benefit from exploring a global threshold and adopting soft labeling rather than the current binary categorization. Furthermore, while our analysis focuses on aggregated LUR curves across the entire dataset, examining the model's behavior on individual utterances may reveal additional insights.

7. Conclusions

In this work, we developed three large-scale speech foundation models for Finnish. They achieved promising results in low-resource ASR (up to 13.9% absolute WERR), and full-scale experiments confirmed their superiority and improved generalizability over existing alternatives, with up to 11.5% absolute WERR on in-domain data and 9.3% on out-of-domain data. Moreover, our best system achieved performance on par with human experts, considerably narrowing the human-AI performance gap.

Our proposed model interpretation technique offers new insights into the behavior of foundation models. Using Layer Utilization Rate (LUR), we observed several interesting trends and found a strong connection between overutilization and low ASR performance. We also found that extremely low or high LUR in the first few layers is a good indicator that the model struggles with new, unseen domains, suggesting that it does not generalize well enough. These results imply that LUR can be utilized to assess whether the model architecture has the potential to generalize to unseen domains or if adding more units and layers could lead to higher performance. Additionally, we observed that LUR curves are heavily influenced by the domain of the pre-training data. This suggests that even after fine-tuning on the same data, the models behave differently, and LUR could help to determine whether the pre-training data were suitable.

While the current experiments use our models only for ASR, in the future, we plan to test them on other speech-based tasks, such as speaker and emotion recognition. At the same time, expanding our analysis to other frameworks remains an interesting task. Lastly, several questions remain open, such as investigating the individual LUR curves per utterance, rather than examining overall trends, or developing a more fine-grained version that goes beyond the two categories used in this study.

8. Acknowledgements

The computational resources were provided by LUMI (project 462000187) and Aalto ScienceIT. We are grateful for the Academy of Finland project funding number 345790 in the ICT 2023 programme's project "Understanding speech and scene with ears and eyes" and the Business Finland project LAREINA under Grant 7817/31/2022.

9. References

[1] A. Moisio *et al.*, "Lahjoita puhetta: a large-scale corpus of spoken finnish with some benchmarks," *Language Resources and Evaluation*, vol. 57, p. 1295–1327, Aug. 2022.

[2] A. Virkkunen, A. Rouhe, N. Phan, and M. Kurimo, "Finnish parliament asr corpus: Analysis, benchmarks and statistics," *Language Resources and Evaluation*, vol. 57, no. 4, pp. 1645–1670, Dec. 2023.

[3] J. Lehečka, J. V. Psutka, L. Smidl, P. Ircing, and J. Psutka, "A comparative analysis of bilingual and trilingual wav2vec models for automatic speech recognition in multilingual oral history archives," in *Interspeech 2024*, 2024, pp. 1285–1289.

[4] P. Szymański *et al.*, "WER we are and WER we think we are," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Association for Computational Linguistics, Nov. 2020, pp. 3290–3295.

[5] P. Bell *et al.*, "Adaptation algorithms for neural network-based speech recognition: An overview," *IEEE Open Journal of Signal Processing*, vol. 2, pp. 33–66, 2021.

[6] D. Hwang *et al.*, "Large-scale asr domain adaptation using self- and semi-supervised learning," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 6627–6631.

[7] C. Wang *et al.*, "VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation," in *Proceedings of the 59th Annual Meeting of the ACL and the 11th ICNLP (Volume 1: Long Papers)*, Aug. 2021, pp. 993–1003.

[8] T. Parcollet *et al.*, "Lebenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech," *Computer Speech & Language*, vol. 86, p. 101622, 2024.

[9] J. Lehečka, J. Švec, A. Prazak, and J. Psutka, "Exploring Capabilities of Monolingual Audio Transformers using Large Datasets in Automatic Speech Recognition of Czech," in *Proc. Interspeech 2022*, 2022, pp. 1831–1835.

[10] M. Malmsten, C. Haffenden, and L. Börjeson, "Hearing voices at the national library – a speech corpus and acoustic model for the swedish language," 2022.

[11] D. Ma, N. Ryant, and M. Liberman, "Probing acoustic representations for phonetic properties," in *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 311–315.

[12] P. Cormac English, J. D. Kelleher, and J. Carson-Berndsen, "Domain-informed probing of wav2vec 2.0 embeddings for phonetic features," in *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, G. Nicolai and E. Chodroff, Eds. Seattle, Washington: Association for Computational Linguistics, Jul. 2022, pp. 83–91.

[13] P. C. English, E. A. Shams, J. D. Kelleher, and J. Carson-Berndsen, "Following the embedding: Identifying transition phenomena in wav2vec 2.0 representations of speech audio," in *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024, pp. 6685–6689.

[14] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 914–921.

[15] A. Pasad, B. Shi, and K. Livescu, "Comparative layer-wise analysis of self-supervised speech models," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[16] A. Saliba, Y. Li, R. Sanabria, and C. Lai, "Layer-wise analysis of self-supervised acoustic word embeddings: A study on speech emotion recognition," in *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSPW)*, 2024, pp. 590–594.

[17] S.-w. Yang *et al.*, "A large-scale evaluation of speech foundation models," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 32, pp. 2884–2899, 2024.

[18] Y. Getman, T. Grosz, and M. Kurimo, "What happens in continued pre-training? analysis of self-supervised speech models with continued pre-training for colloquial finnish asr," in *Interspeech 2024*, 2024, pp. 5043–5047.

[19] A. Barredo Arrieta *et al.*, "Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai," *Information Fusion*, vol. 58, pp. 82–115, 2020.

[20] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ser. ICML'17. JMLR.org, 2017, p. 3319–3328.

[21] H. Bredin, "pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe," in *INTERSPEECH 2023*, 2023, pp. 1983–1987.

[22] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," *Advances in neural information processing systems*, vol. 33, pp. 12 449–12 460, 2020.

[23] A. Babu *et al.*, "Xls-r: Self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022*, 2022, pp. 2278–2282.

[24] T. Grósz, A. Virkkunen, D. Porjazovski, and M. Kurimo, "Discovering Relevant Sub-spaces of BERT, Wav2Vec 2.0, ELECTRA and ViT Embeddings for Humor and Mimicked Emotion Recognition with Integrated Gradients," in *Proceedings of the 4th Multimodal Sentiment Analysis Challenge and Workshop*. ACM, 2023.

[25] N. Kokhlikyan *et al.*, "Captum: A unified and generic model interpretability library for pytorch," 2020.

[26] M. Ott *et al.*, "fairseq: A fast, extensible toolkit for sequence modeling," in *Proceedings of NAACL-HLT 2019: Demonstrations*, 2019.

[27] "Lumi supercomputer," https://www.lumi-supercomputer.eu, accessed: 2024-08-20.

[28] R. Ardila *et al.*, "Common voice: A massively-multilingual speech corpus," in *Proceedings of the Twelfth Language Resources and Evaluation Conference*, N. Calzolari *et al.*, Eds. Marseille, France: European Language Resources Association, May 2020, pp. 4218–4222.

[29] A. Conneau *et al.*, "Fleurs: Few-shot learning evaluation of universal representations of speech," in *2022 IEEE Spoken Language Technology Workshop (SLT)*, 2023, pp. 798–805.

[30] T. Wolf *et al.*, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on EMNLP: System Demonstrations*. Online: Association for Computational Linguistics, Oct. 2020, pp. 38–45.

[31] L. Ferrer and P. Riera, "Confidence Intervals for evaluation in machine learning." [Online]. Available: https://github.com/luferrer/ConfidenceIntervals