Towards large-scale speech foundation models for a low-resource minority language

Yaroslav Getman
Department of lnformation and Communications Engineering
Aalto University
Finland
yaroslav.getman@aalto.fi

Tamás Grósz
Department of lnformation and Communications Engineering
Aalto University
Finland
tamas.grosz@aalto.fi

Katri Hiovain-Asikainen
UiT The Arctic University of Norway
katri.hiovain-asikainen@uit.no

Tommi Lehtonen
Finnish National Audiovisual Institute (KAVI)
Finland
tommi.lehtonen@kavi.fi

Mikko Kurimo
Department of lnformation and Communications Engineering
Aalto University
Finland
mikko.kurimo@aalto.fi

Abstract

Modern ASR systems require massive amounts of training data. While ASR training data for most languages are scarce and expensive to transcribe, a practical solution is to collect huge amounts of raw untranscribed speech and pre-train the ASR model in a self-supervised manner. Unfortunately, for many low-resource minority languages, even untranscribed speech data are scarce. In this paper, we propose a solution for the Northern Sámi language with 22,400 hours of speech extracted from the Finnish radio and television archives. We evaluated the model performance with different decoding algorithms and examined the models' internal behavior with interpretation-based techniques.

1 Introduction

Self-Supervised Learning (SSL) has caused a paradigm shift in Automatic Speech Recognition (ASR), enabling the development of highly accurate End-to-End models even with a limited amount of data. Low-resource languages also benefited from this advancement, as models pre-trained on other languages proved to be a good foundation for the development of ASR models using small supervised corpora (Bogdanoski et al., 2023; Gilles et al., 2023). Northern Sámi, a language spoken by only about 20,000 people has also seen rapid advancements in speech technology (Hiovain-Asikainen and De la Rosa, 2023; Getman et al., 2024a).

While fine-tuning speech foundation models such as wav2vec 2.0 (Baevski et al., 2020) can now be considered standard procedure, choosing the right pre-trained system is still very critical. Several works have reported that monolingual pre-training tends to produce the best foundation (Evain et al., 2021; Lehečka et al., 2024; Parcollet et al., 2024), which could be impossible without access to large speech-only corpora. Alternatively, continuing the pre-training of an existing model could adapt it to new languages (Javed et al., 2022). In this work, we build speech foundation models for Northern Sámi with about 22,400 hours of speech from radio broadcasts, which puts them on par with most publicly available monolingual speech foundation models for high-resource languages (Evain et al., 2021; Wang et al., 2021; Javed et al., 2022; Malmsten et al., 2022; Getman et al., 2024b; Parcollet et al., 2024; Sawada et al., 2024).

In the past, various training methods have been explored for wav2vec 2.0. Still, its inference is

Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pages 192–200

most commonly done via a greedy decoding algorithm. Here, we explore whether a more advanced technique called prefix beam search (Hannun et al., 2014) could lead to better results. The main issue with the standard greedy algorithm stems from the blank symbol, which usually receives a considerable portion of the probability mass (Jung et al., 2022), thus leading to spiky outputs and many deletion errors. To avoid this unwanted effect, prefix beam search merges multiple paths that would result in the same output, lessening the suppression effect of the blank output. While this technique was originally proposed to be used with recurrent models, its variants have been successfully utilized with large SSL models (Jung et al., 2022) and encoder-decoder-based architectures (Zhao et al., 2024) too, albeit those works also employ an external LM during the decoding procedure. In contrast, we only utilize prefix beam search to decode the wav2vec 2.0 model without any LM parts, as low-resource languages often lack in terms of text data too, which prevents the development of a good LM.

Besides the training and decoding algorithms, we also take a closer look at our models' mistakes and propose a new interpretation-based solution to learn more about the reasons for the misrecognition. One of our main observations revealed systematic, repeating mistakes, which we hypothesized were due to the dominance of the Internal LM developed by the model during the finetuning phase (Zeyer et al., 2021a). To validate this hypothesis, we utilized the Integrated Gradients (IG) technique (Sundararajan et al., 2017) to investigate whether the model behaves differently when it predicts various characters. Our experiments revealed that several characters which caused the problems were predominantly outputted by using mainly the long-term information embeddings while ignoring the current acoustic information. Furthermore, we have found that the model dedicated considerably more neurons towards detecting the rare Sámi-specific characters compared to the common Latin characters.

In summary, in this paper, we made the following contributions:

- Developed the first Northern Sámi speech foundation models [1].
- Compared the greedy decoding algorithm with the prefix beam search algorithm without any LM component.
- Proposed a model interpretation technique to investigate why the model makes certain mistakes.

2 Methods

2.1 Continued Pre-Training

While standard pre-training of wav2vec 2.0 implies random initialization of the model weights, another training option is utilizing weights of an existing foundation model from a closely related language(s). Getman et al. (2024a) has demonstrated that continued pre-training on a small, 100-hour dataset can improve the downstream out-of-domain ASR performance. In this work, we take a step further and analyze whether this technique is useful even when a sufficient amount of unlabeled in-domain data is available.

Continued pre-training differs from pre-training from scratch only during the model initialization phase; otherwise, it follows the same standard training pipeline. A side effect of this approach is catastrophic forgetting (McCloskey and Cohen, 1989), which hinders the models' performance on language(s) they have been originally pre-trained on (Qian et al., 2024). However, one of the goals of this work is to develop monolingual foundations for a low-resource minority language rather than expand the mono- and multilingual models' capabilities to a new language.

2.2 Prefix Beam Search

End-to-end ASR models like wav2vec 2.0 are often trained with the Connectionist temporal classification (CTC) algorithm in the finetuning phase (Graves et al., 2006). While CTC offers a convenient way of training, the resulting models are well-known to suffer from various problems; namely, the blank label introduced by CTC usually obtains very high probabilities dominating the sequence of outputted symbols, and non-blank outputs display a peaky behavior (Zeyer et al., 2021b). These problems together mean that CTC-trained models often have high deletion errors, as the blank label could easily suppress the emission of actual characters, especially when the model is uncertain.

Prefix Beamsearch (Hannun et al., 2014) offers an alternative to the standard greedy decod-

[1] https://huggingface.co/collections/GetmanY1/wav2vec2-sami-22k-66ead12fe465d6302b63d11b

ing algorithm by considering multiple paths that would result in the same output and combining the probabilities of these paths to gain a more accurate estimate of character emission probabilities. For example, if we consider a short window of 4 timesteps in which the model should recognize the character "a", then the greedy decoding would require that the output unit linked to "a" would get the maximum probability at least in one frame. In many cases, this assumption is not true. Thus, the character is deleted if the probability of the blank ($\varnothing$) is high. In the beam search algorithm, all possible combinations of $\varnothing$ and "a" are considered, and the probabilities of these paths (e.g. $\varnothing\varnothing a\varnothing$, or $\varnothing aa\varnothing$, or $\varnothing\varnothing aa$, etc.) are added together, often surpassing the probability of purely $\varnothing$ output, preventing the character deletion problem.

The algorithm was originally proposed for recurrent models, and RNN-T architectures, but here we demonstrate that it is applicable even with wav2vec 2.0 models, without any LM. In practice, we fix all the LM probabilities as 1 and feed the logit values of wav2vec 2.0 after a softmax layer to the decoding algorithm.

2.3 IG-based error analysis

For a long time, large foundation models, like any other deep neural network, were considered a black box. With the advancement made in the field of model explainability (Schwalbe and Finzel, 2021), it is now possible to peak inside these huge models and investigate their internal functions. In this work, we selected the technique called Integrated Gradients (IG) (Sundararajan et al., 2017) to learn more about the internal representations of our systems. IG belongs to the family of gradient-based posthoc interpretation tools, meaning that no modifications of the training algorithm or the model architecture are needed to gain insight. In essence, IG estimates the gradients of the relevant output units with respect to certain hidden neurons, and these values are called attributions. In Grósz et al. (2023), it was demonstrated that IG can be used to filter out the irrelevant neurons of various foundation models, without any significant performance loss. Inspired by these findings, here we employed IG to unveil how our models predict certain characters.

Our primary goal was to understand when the model makes decisions mainly based on acoustic information, and when the Internal LM (ILM) (Zeyer et al., 2021a) becomes dominant. Several techniques have already been proposed to estimate the ILM developed during supervised training. In Zeyer et al. (2021a); Chen et al. (2023), the authors suggest masking out the encoder (acoustic) output to find the ILM scores or employing the so-called density ratio method. Unfortunately, these techniques are not applicable in our case as our model does not have a decoder part, and it is not autoregressive, thus we developed an alternative IG-based solution.

In our experiments, we choose to focus on two specific layers of the wav2vec 2.0 model; namely the feature embeddings of the CNN component, which can be considered as acoustic features, and the convolutional positional embedding layer's output, where temporal information is introduced to the model. Using IG, we estimated the attributions of each neuron in these two layers per output units. Here we used the predicted (most probable) output at each timestep to estimate the attributions. Next, to approximate the importance of each layer, we calculated the sum of the absolute attributions of neurons inside the two layers. Our motivation for using the absolute values was simple; we did not want to lose valuable information if some neurons had both large negative and positive attributions at different times. Lastly, once the overall attribution of the two layers' was known, the attribution ratio was calculated by dividing the positional embedding layer's attribution by the feature embedding layer's. In this context, an attribution ratio of 1 means that the positional embedding layer has the exact same information as the feature embeddings (i.e. it has no extra influence on the outputs), while a ratio of 2 means that the new temporal information introduced by the positional embedding layer is equally important compared to the acoustic one. Naturally, a ratio above 2 implies that the temporal information is valued more than the acoustic features, which is a sign of the ILM dictating the final output.

3 Data

For pre-training the Sámi models, we extracted 35,614 hours of radio broadcasts of Yle Saamen Radio. The broadcasts have been originally recorded by the Radio and Television Archive (RTVA) since 2009 and provided for research by the Finnish National Audiovisual Institute (KAVI).

Since the raw dataset also contained a considerable amount of non-speech events, including music and silence, we pre-processed it with a neural voice activity detector (VAD) (Bredin, 2023). After that, continuous speech segments longer than 30 seconds were split into shorter utterances. The final size of pre-training data was 22,415 hours, meaning that nearly 37% of the audio was recognized as non-speech.

For the ASR fine-tuning, we used the Sámi Parliament data [2] featuring about 20 hours of transcribed speech. For testing the ASR models, 1 hour of out-of-domain read-aloud and spontaneous speech of varying audio quality was used.

4 Experiments

We pre-trained the foundation models with the Fairseq toolkit (Ott et al., 2019). Pre-training was done on 512 GPUs of the LUMI supercomputer [3] for 125,000 steps (approx. 115 epochs) for the Base models (95M parameters) and 167,000 steps (approx. 100 epochs) for the Large ones (317M parameters). The models were then fine-tuned on the Sámi Parliament data for 60 epochs with Huggingface Transformers (Wolf et al., 2020). In continued pre-training, we adapted models originally pre-trained on European Parliament plenary session recordings (Wang et al., 2021). The Base model was a monolingual Finnish foundation, while the Large one also included speech from two other Uralic languages (Hungarian and Estonian).

We evaluated the models with the standard ASR performance metrics such as word and character error rate (WER and CER) and compared them to existing ASR solutions, including Whisper (Radford et al., 2023) fine-tuned on 34 hours of spontaneous Northern Sámi (Hiovain-Asikainen and De la Rosa, 2023) and XLS-R (Babu et al., 2022) first fine-tuned on high-resource Finnish data and then adapted to Northern Sámi with the Sámi Parliament data (Getman et al., 2024a).

Table 1 summarizes the ASR results. Compared to the previously developed solutions, the Base-sized models provided lower WER but higher CER. In contrast, when switching to the Large models, more considerable improvements can be observed.

Next, we performed statistical significance tests on both the word and character levels using the Matched Pair Sentence Segment approach. To run the tests, we employed the SCTK toolkit [4]. Looking at the models with continued pre-training, models of both sizes gave significantly ($p \leqslant 0.001$) lower CER compared to pre-training from scratch, but only the Large one significantly ($p \leqslant 0.05$) outperformed its counterpart pre-trained from scratch on the word level.

Switching from greedy decoding to prefix beam search further improved the CER. On the word level, however, a significant improvement can be observed only for the Large model pre-trained from scratch, while it insignificantly changed the error rate in either direction for the rest of the models. A more detailed analysis of the results revealed that the prefix beam search always increased the number of substitutions and insertions but decreased the number of deletions compared to greedy decoding.

Overall, the best results were obtained by continued pre-training of the Large model. It gave a noticeable improvement on a character level over pre-training on the same data from scratch (14% relative CER reduction), which may suggest that continued pre-training allowed the model to benefit from acoustic patterns learned from other languages and combine them with the newly learned acoustic information of the target language. On the other hand, minor changes in the WER and the distribution of error rates in Figure 1 may indicate that the gained language knowledge was still not sufficient enough to properly recognize complete words.

5 Analysis of the results

To better understand how our best model (*Large-22K CPT + Prefix Beam Search*) works, and why it makes certain mistakes, we first inspected the character-level confusion matrix on the test data, see Figure 2. Overall, most characters could be recognized with relatively good accuracy, and only a few rare characters like å, ä, x, ö have extremely low recognition rates. While these mistakes can be explained by the lack of training data, we also noticed other systematic problems on the word level. One such issue was related to the word "na" (in English: "well"), which was quite common in the training data. Interestingly, in the test set other similar words, like "ni" and "no" were almost al-

[2] https://sametinget.kommunetv.no/archive

[3] https://www.lumi-supercomputer.eu/

[4] https://github.com/usnistgov/SCTK

System	WER, %	CER, %
XLS-R EFT (Getman et al., 2024a)	47.70	15.15
Whisper (Hiovain-Asikainen and De la Rosa, 2023)	43.15	14.05
Base-22K	43.07	16.50
Base-22K + Prefix Beam Search	43.12	16.20***
Base-22K CPT	43.04	15.76
Base-22K CPT + Prefix Beam Search	42.74	15.51***
Large-22K	33.32	12.76
Large-22K + Prefix Beam Search	32.94**	12.51***
Large-22K CPT	32.28	10.83
Large-22K CPT + Prefix Beam Search	32.29	10.76**

Table 1: WER and CER on the 1-hour out-domain test set. EFT = extended fine-tuning; CPT = continued pre-training. Statistically significant improvements of the prefix beam search over the greedy decoding are marked ***$p \leqslant 0.001$, **$p \leqslant 0.01$

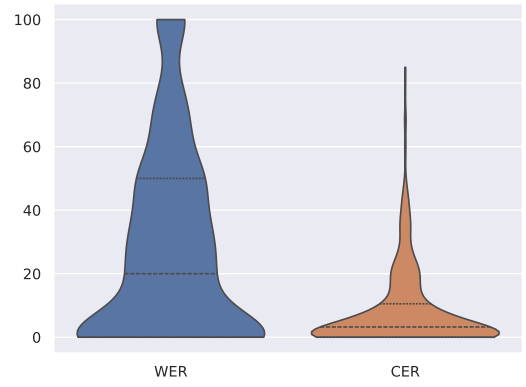


Figure 1: The distribution of the WER and CER per utterance of our best model (Large-22K CPT + Prefix Beam Search) on the test set. Utterances with more than 100% error rates were pooled together for the visualization.

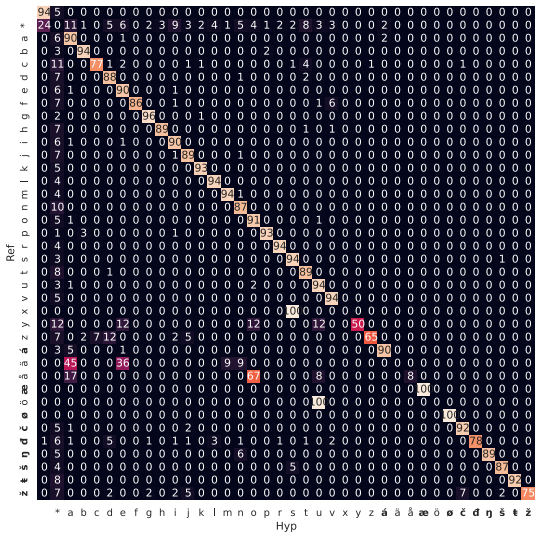
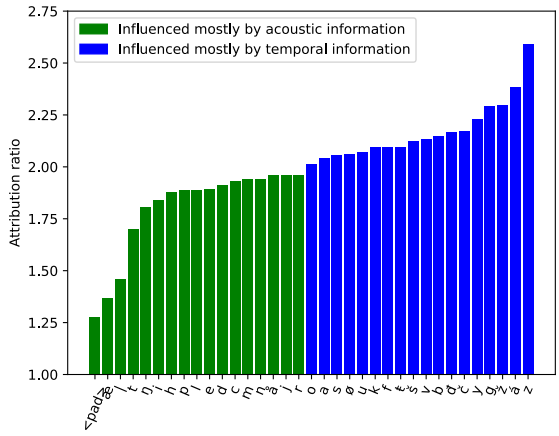


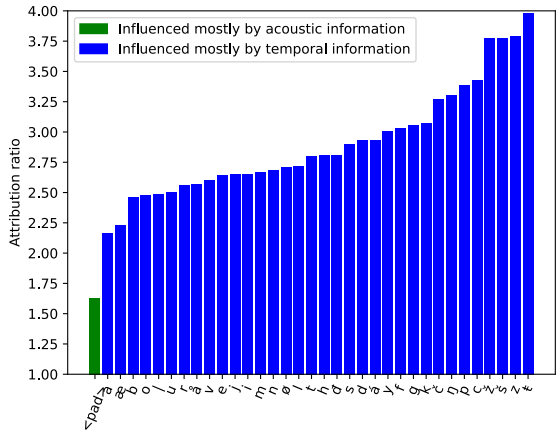
Figure 2: The character confusion matrix of our best model (Large-22K CPT + Prefix Beam Search) on the test set.

ways replaced by the word "na", which implied that the model developed a strong internal LM, which forced it to predict the character "a" after the letter "n", especially at the beginning of the sentence, when the model has limited context.

To validate this hypothesis about the internal LM, we employed our proposed solution to better understand why our best wav2vec 2.0 made certain mistakes. Our first observation was that the overall attribution ratio on the whole set was above 1.3, proving that the temporal information introduced by the positional embedding layer was indeed utilized by the model, but acoustic features were valued considerably more. Furthermore, we saw that the blank and the word boundary symbols have the lowest ratios, signaling that they were predicted mostly based on acoustic information. Looking at the actual characters, we saw that the average ratio was approximately 2, suggesting that, on average, the temporal information introduced in the positional embedding layer was as useful as the acoustic features extracted by the CNN component.



(a) Continued pre-training.



(b) Pre-training from scratch.

Figure 3: The attribution ratios between the positional and features embeddings per character.

After a closer look at the ratios per character (see Figure 3a), we identified two groups; in the first one, the ratio was below 2, suggesting that these were predicted mainly using acoustic features. This group includes characters such as "h", "i", "n", etc. On the other hand, we can see several characters, including "a", which were primarily predicted by the influence of the internal LM. These results imply that for some characters the acoustic component of the model was not good enough, and it would benefit from seeing additional training material with more diverse textual content in order to force the model to rely more on the acoustic information.

Next, we investigated the counterpart of the best model, trained from scratch (Large-22K), see Figure 3b. This model demonstrated a quite different behavior: all tokens except the blank label had an attribution ratio above 2, meaning that the system's output was determined mostly by the temporal information added by the positional embeddings. The average attribution ratio for non-blank characters was 2.7, signaling that the acoustic component had a considerably smaller attribution towards the output than the internal LM. Considering that the model was pre-trained only with a relatively small dataset, we can conclude that the acoustic component produced by the continued pre-training is more appropriate and extracts more relevant information. The purely Northern Sámi model's overreliance on temporal information indicates that it most probably obtained most of its knowledge by simply memorizing parts of the training transcripts during the fine-tuning phase, as large models are prone to do so (Huang et al., 2022; Wang et al., 2024). Validating this theory is out of the scope of this paper, but remains an important future task.

Lastly, we also investigated individual neurons in the two selected layers. Here, we aimed to find out which character needed the most actively contributing neurons. We looked at each neuron's attribution values per character. First, we calculated the average and standard deviation of the attributions in each layer. Our first observation at this stage was that the majority of the neurons had an attribution close to the mean (which was approximately 0 in all cases), and only a few neurons displayed large attributions similar to the findings of (Grósz et al., 2023). Based on these observations, we decided to separate the neurons into two groups; the highly contributing ones, whose accumulated attribution was farther than one standard deviation from the mean, and the rest categorized as low-contributing.

Figure 4 illustrates the amount of highly attributing neurons in each investigated layer of the best model. The first observation is that common characters like "r", "b" and "k" required only a few dedicated neurons, while special Sámi characters like "ŧ" and "æ" were predicted based on a large number of neurons. In general, many latin characters required less than a 100 highly contributing neurons, while many Sámi charaters needed more units. This implies that the acoustic features of the CPT model were quite good for most Latin characters that were well represented in the original pre-training corpus, while some ("d", "c" and "t") required more units, perhaps due to non-standard

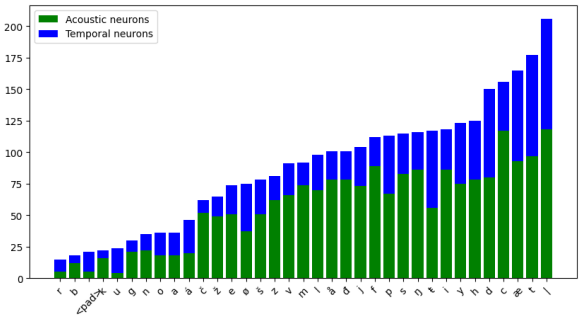


Figure 4: Number of highly attributing neurons in the best model. Acoustic neurons refer to units in the feature embedding layer, while Temporal ones can be found in the positional embedding layer.

pronunciation. Additionally, we can see that the model dedicated a larger portion of neurons to the Sámi specific outputs, implying that despite the language adaptation via CPT and finetuning, it still has difficulties recognizing them well.

6 Limitations

While experimental results suggest that prefix beam search is beneficial on the character level, its WERs proved to be quite similar to the greedy decoding algorithm's. As the lower CER suggests better quality output, testing its readability by humans and comparing it to the greedy alternative remains an important future task. Additionally, we should mention that here, we utilized the prefix search without any modifications, but it might benefit from adjustments in terms of hyperparameters and vocabulary usage of wav2vec 2.0, especially regarding the word separator symbol.

While our model interpretation experiments have revealed interesting facts about the internal functions of the models, they should be rigorously tested and validated. On the one hand, interpretation techniques are known to be fragile (Ghorbani et al., 2019). Thus, our experiments should be repeated with other attribution estimation methods to ensure that our observations hold. Furthermore, we made several simplifications in this work, including the decision to accumulate the attributions over time, thus ignoring their changes in different contexts. In the future, we intend to investigate how the attributions' trajectories change over time and in different contexts to gain a deeper understanding of when temporal information is valued more than acoustic information. Lastly, all of our findings should be validated by the use of a reliable ILM estimation method. Unfortunately, currently, no such technique is available for non-autoregressive models.

7 Conclusions

In this work, we presented the first speech foundation models for Northern Sámi. In addition to standard greedy decoding, we tested prefix beam search, which showed a slight improvement in terms of CER by reducing the number of deletions. Although continued pre-training of a multilingual foundation did not bring a considerable improvement in downstream ASR performance compared to pre-training from scratch, deeper IG-based analysis demonstrated differences in the internal behavior of these two models and revealed that the one pre-trained from scratch was heavily influenced by the temporal information (internal LM), while its counterpart with continued pre-training relied more on its acoustic component when predicting certain characters.

References

Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick von Platen, Yatharth Saraf, Juan Pino, Alexei Baevski, Alexis Conneau, and Michael Auli. 2022. Xls-r: Self-supervised cross-lingual speech representation learning at scale. In *Interspeech 2022*, pages 2278–2282.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460. Curran Associates, Inc.

Konstantin Bogdanoski, Kostadin Mishev, Monika Simjanoska, and Dimitar Trajanov. 2023. Exploring asr models in low-resource languages: Use-case the macedonian language. In *Deep Learning Theory and Applications*, pages 254–268, Cham. Springer Nature Switzerland.

Hervé Bredin. 2023. pyannote.audio 2.1 speaker diarization pipeline: principle, benchmark, and recipe. In *INTERSPEECH 2023*, pages 1983–1987.

Zhipeng Chen, Haihua Xu, Yerbolat Khassanov, Yi He, Lu Lu, Zejun Ma, and Ji Wu. 2023. Knowledge distillation approach for efficient internal language model estimation. In *INTERSPEECH 2023*, pages 1339–1343.

Solène Evain, Ha Nguyen, Hang Le, Marcely Zanon Boito, Salima Mdhaffar, Sina Alisamir, Ziyi Tong,

Natalia Tomashenko, Marco Dinarelli, Titouan Parcollet, Alexandre Allauzen, Yannick Estève, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2021. LeBenchmark: A Reproducible Framework for Assessing Self-Supervised Representation Learning from Speech. In *Proc. Interspeech 2021*, pages 1439–1443.

Yaroslav Getman, Tamas Grosz, Katri Hiovain-Asikainen, and Mikko Kurimo. 2024a. Exploring adaptation techniques of large speech foundation models for low-resource ASR: a case study on Northern Sámi. In *Interspeech 2024*, pages 2539–2543.

Yaroslav Getman, Tamas Grosz, and Mikko Kurimo. 2024b. What happens in continued pre-training? analysis of self-supervised speech models with continued pre-training for colloquial finnish asr. In *Interspeech 2024*, pages 5043–5047.

Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of neural networks is fragile. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):3681–3688.

Peter Gilles, Léopold Edem Ayité Hillah, and Nina Hosseini Kivanani. 2023. Asrlux: Automatic speech recognition for the low-resource language luxembourgish. In *Proceedings of the 20th International Congress of Phonetic Sciences*. Guarant International.

Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber. 2006. Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. *Proceedings of the 23rd International Conference on Machine Learning*, page 369–376.

Tamás Grósz, Anja Virkkunen, Dejan Porjazovski, and Mikko Kurimo. 2023. 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: Mimicked Emotions, Humour and Personalisation (MuSe '23)*. ACM.

Awni Y. Hannun, Andrew L. Maas, Daniel Jurafsky, and Andrew Y. Ng. 2014. First-Pass Large Vocabulary Continuous Speech Recognition using Bi-Directional Recurrent DNNs.

Katri Hiovain-Asikainen and Javier De la Rosa. 2023. Developing tts and asr for lule and north sámi languages. In *2nd Annual Meeting of the ELRA/ISCA SIG on Under-resourced Languages (SIGUL 2023)*, pages 48–52.

W. Ronny Huang, Steve Chien, Om Dipakbhai Thakkar, and Rajiv Mathews. 2022. Detecting unintended memorization in language-model-fused asr. In *Interspeech 2022*, pages 2808–2812.

Tahir Javed, Sumanth Doddapaneni, Abhigyan Raman, Kaushal Santosh Bhogale, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2022. Towards building asr systems for the next billion users. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10813–10821.

Minkyu Jung, Ohhyeok Kwon, Seung Byum Seo, and Soonshin Seo. 2022. Blank collapse: Compressing ctc emission for the faster decoding. *ArXiv*, abs/2210.17017.

Jan Lehečka, Josef V. Psutka, Lubos Smidl, Pavel Ircing, and Josef Psutka. 2024. A comparative analysis of bilingual and trilingual wav2vec models for automatic speech recognition in multilingual oral history archives. In *Interspeech 2024*, pages 1285–1289.

Martin Malmsten, Chris Haffenden, and Love Börjeson. 2022. Hearing voices at the national library – a speech corpus and acoustic model for the swedish language.

Michael McCloskey and Neal J. Cohen. 1989. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of NAACL-HLT 2019: Demonstrations*.

Titouan Parcollet, Ha Nguyen, Solène Evain, Marcely Zanon Boito, Adrien Pupier, Salima Mdhaffar, Hang Le, Sina Alisamir, Natalia Tomashenko, Marco Dinarelli, Shucong Zhang, Alexandre Allauzen, Maximin Coavoux, Yannick Estève, Mickael Rouvier, Jerôme Goulian, Benjamin Lecouteux, François Portet, Solange Rossato, Fabien Ringeval, Didier Schwab, and Laurent Besacier. 2024. Lebenchmark 2.0: A standardized, replicable and enhanced framework for self-supervised representations of french speech. *Computer Speech & Language*, 86:101622.

Mengjie Qian, Siyuan Tang, Rao Ma, Kate Knill, and M.J.F. Gales. 2024. Learn and don't forget: Adding a new language to asr foundation models. In *Interspeech 2024*, pages 2544–2548.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Kei Sawada, Tianyu Zhao, Makoto Shing, Kentaro Mitsui, Akio Kaga, Yukiya Hono, Toshiaki Wakatsuki, and Koh Mitsuda. 2024. Release of pre-trained

models for the Japanese language. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13898–13905, Torino, Italia. ELRA and ICCL.

Gesina Schwalbe and Bettina Finzel. 2021. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, pages 1–59.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 3319–3328. JMLR.org.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. 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 IC-NLP (Volume 1: Long Papers)*, pages 993–1003.

Lun Wang, Om Thakkar, Zhong Meng, Nicole Rafidi, Rohit Prabhavalkar, and Arun Narayanan. 2024. Efficiently train asr models that memorize less and perform better with per-core clipping. In *Interspeech 2024*, pages 1320–1324.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on EMNLP: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Albert Zeyer, Andr'e Merboldt, Wilfried Michel, Ralf Schluter, and Hermann Ney. 2021a. Librispeech transducer model with internal language model prior correction. In *Interspeech*.

Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2021b. Why does CTC result in peaky behavior? *CoRR*, abs/2105.14849.

Zeyu Zhao, Peter Bell, and Ondřej Klejch. 2024. Exploring Dominant Paths in CTC-Like ASR Models: Unraveling the Effectiveness of Viterbi Decoding. In *2024 IEEE International Conference on Acoustics, Speech, and Signal Processing Workshops (ICASSP)*, pages 868–872.