INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

Combining Multilingual Resources and Models to Develop State-of-the-Art E2E ASR for Swedish

Lukas Mateju, Jan Nouza, Petr Cerva, Jindrich Zdansky, Frantisek Kynych

Faculty of Mechatronics, Informatics and Interdisciplinary Studies,
Technical University of Liberec, Studentska 2, 461 17 Liberec, Czech Republic

lukas.mateju@tul.cz

Abstract

In terms of automatic speech recognition (ASR), Swedish belongs to the group of less-resourced languages, as publicly available training data is limited to a few hundred hours of mostly read speech. To acquire larger amounts of more realistic data, we investigate the existing multilingual approaches, and also propose two new ones, which combine Swedish with previously created Norwegian data and models. We use them for efficient automatic harvesting of spoken Swedish from broadcast, parliament, YouTube, and audiobook archives. The combined models significantly speed up the harvesting process and improve the final Swedish end-to-end (E2E) ASR system. We evaluate it on datasets covering various applications and domains; they provide performance better than the state-of-the-art commercial cloud services. We have made all of our test datasets publicly available for future comparative experiments.

Index Terms: end-to-end speech recognition, transfer learning, multilingual training, Scandinavian languages, Swedish

1. Introduction

Swedish (SWE), together with Norwegian (NO) and Danish, belongs to the North-Germanic language branch. It is the most utilized of these languages, spoken by 9.2M people. (The others have about 5M speakers each.) Research in the field of ASR for Scandinavian languages was active mainly in the 1990s while the two following decades brought just modest progress. There were several reasons for that, the main being limited speech resources. Moreover, the traditional ASR approach requires extensive vocabularies (due to compound words and suffixed articles) and many pronunciation variants (due to dialects).

The advent of modern E2E systems [1] has partly eliminated such issues; and this progress has motivated new research activities. Several multilingual systems (e.g., [2, 3, 4]), also covering Swedish, have been developed, but they have mainly been aimed at experimenting with joint models for multiple languages, rather than targeting the best possible monolingual performance. Some recently completed MSc theses have focused on adopting the E2E approaches solely to Swedish, and at least two papers have been published in this area since 2020 [5, 6]. Their authors tried to utilize existing E2E frameworks (namely unsupervised wav2vec 2.0 [7]) and available models together with public datasets (NST and Common Voice – see Sec. 4.1) to tune the models to be viable specifically for Swedish ASR. Results obtained on these two sets look very promising, with a word error rate (WER) below 10%. However, experiments with independent and more realistic data yield notably worse performance. This observation confirms the well-known fact that E2E systems require much larger training resources covering different speaking styles, various environments, and diverse topics.

The easiest way to collect more training data is to use public sources such as broadcast, parliament, or YouTube archives that offer both audio and text. Topic and lexicon diversity can be enlarged by utilizing audiobooks (along with corresponding e-books), which are available in large quantities, even though not free of charge. As none of these sources provides data that could directly be used for training, we need tools that perform three basic tasks: a) prepare audio and text data for alignment; b) detect parts that match; and c) split them into files eligible for training. This can (almost) automatically be done by first employing an existing ASR system that converts audio files into text before the three tasks are consecutively executed. The data harvesting process is iterative, starting with an initialized (bootstrapped) system, which is repeatedly retrained on increasing amounts of data. However, for less-resourced languages, this iterative process can be rather slow and inefficient; we therefore investigate methods to boost it by leveraging data and models (or their parts) already available for other languages.

In this paper, the target language is Swedish, and we present and compare several multilingual techniques, such as transfer learning and multilingual training, which allow us to utilize previously created Norwegian data and E2E models to collect more than 1,000 hours of realistic training data and make an ASR model that performs well in various practical applications.

2. Related work

Due to the already mentioned data hunger of E2E systems, the importance of the bootstrapping phase is even greater now, particularly for low-resourced languages. In recent years, several techniques, especially transfer learning [8, 9] and multilingual training [10, 11], have proven to be successful in bootstrapping E2E models. In transfer learning, the model for the new (target) language is initialized from an already existing ASR model of a high-resourced language, while a joint model is trained on data from multiple languages (including the target one) within multilingual training. These techniques can be combined [12, 2] and also used in conjunction with other methods, such as semi-supervised learning [13], data augmentation [14], and text-to-speech utilization [15].

Transfer learning has recently been applied to bootstrapping different E2E architectures. A pre-trained multilingual model was used to initialize several low-resource languages in a hybrid connectionist temporal classification [16] and attention-based encoder-decoder [17] (CTC/AED) model [18, 12]. For the recurrent neural networks transducer (RNN-T) [19] framework, four different variations of transfer learning were studied in [20] and a combination of transfer learning with multilingual training, text-to-text mapping, and synthetic audio techniques in [21].

10.21437/Interspeech.2023-737

The focus of the E2E multilingual training lies in two interconnected directions – training a massive model capable of transcribing as many languages as possible [2, 3, 22], and improving the performance of low-resourced languages by utilizing high-resourced (related) ones [23, 24]. For the latter direction, different E2E architectures, such as hybrid CTC/AED [18], and RNN-T [21], have been exploited. Furthermore, the overall performance can be improved by providing language identity information that guides the system to produce better transcriptions of the target language [25]. Commonly, this is done by appending a one-hot vector [26], or a learned language embedding [27]. Moreover, a joint language-independent multilingual ASR with language identification has been proposed in [28, 12]. Multi-headed models (per language group) have also been explored [2]. Lastly, the issue of data imbalance between high- and low-resource languages has also been studied [29].

Lately, self-supervised learning, a technique extracting general data representations from unlabeled data, has produced many excellent results even when fine-tuned on severely limited amounts of data, e.g., wav2vec 2.0 [7] or w2v-BERT [30]. In [31], the authors built on the wav2vec 2.0 and proposed cross-lingual speech representation (XLSR). XLS-R [32] further extended the previous works. Joint unsupervised and supervised training for multilingual ASR was proposed in [33].

3. E2E architecture and its modifications

Within this work, the ESPNet platform [34] empowers the adopted end-to-end architecture corresponding to the joint combination of CTC and AED. Our model thus comprises three parts: a shared encoder represented by a conformer [35] and two decoders: CTC-based and attention-based, using a CTC weighting factor equal to 0.3.

The shared encoder is preceded by two sub-sampling convolutional layers (kernel size 3×3 and stride 2), and it includes 12 blocks, each having eight attention heads. The CTC decoder consists of a linear layer, which transforms the encoder output to the CTC activation. The attention decoder is a transformer, and it contains six blocks, each with eight attention heads. In each model block, the dimension of attention is set to 512, and the number of units of the position-wise feed-forward layer is 2,048. In total, the entire model contains 136M of parameters.

The input speech is parameterized to 80-dimensional Mel-spectral filter banks (25 ms long), and SpecAugment [14] is applied on the fly to augment training data (with a length limited to 25 s). The model is trained for 120 epochs using the Adam optimizer with a batch size of 20, and the final model is obtained by averaging 30 epochs with the lowest loss on a 10-hour dev set. For decoding, the CTC prefix beam search algorithm is used.

It has been shown [36] that a slight boost to E2E ASR can be achieved if the output symbols are not just single letters but word fragments derived from the most frequent words. In our case, the SentencePiece toolkit [37] is used to get the 5,000 most frequent tokens from input speech lower-case annotations.

Moreover, the encoder can be initialized from an already existing model (e.g., of a high-resource language) via transfer learning [18]. A comparison between the non-initialized (*mono*) and initialized (*init*) approaches is depicted in Fig. 1a) and 1b).

3.1. Adopted and proposed multilingual modifications

We explore four different variations of multilingual training. In the first case (denoted *joint*), the training datasets of multiple languages are combined into a single training set, and the model is simply trained as a monolingual one, as shown in Fig. 1c). The number of model parameters remains the same (i.e., 136M).

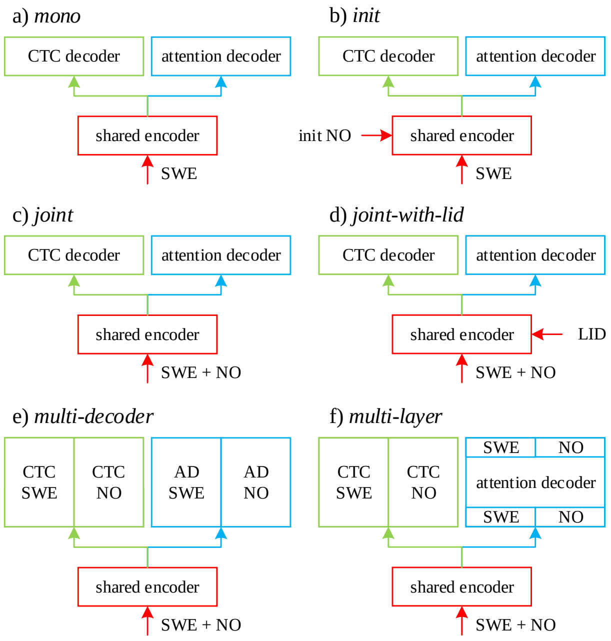

Figure 1: *Adopted architecture and its modifications applied to Swedish as a target language and Norwegian as a support one.*

The second variation (*joint-with-lid*) outlined in Fig. 1d) extends the previous modification. In this case, the adopted model is guided by a one-hot language identity vector [26] appended to the Mel-filter banks during both the training and decoding.

The third variant is further denoted as *multi-decoder*. This proposed model (similar to [2]) comprises a pair of language-specific decoders for every language, as depicted in Fig. 1e). It thus has 44M language-specific parameters for every language and 92M parameters shared among all of the languages.

The last modification (*multi-layer*) we propose employs a language-specific CTC decoder for each language and a shared attention decoder with language-specific layers, as highlighted in Fig. 1f). These layers correspond to input embeddings and output linear layers of all blocks. The main advantage is that the parameters of all attention heads can thus be trained using all the multilingual data. This model only has 13M language-specific parameters per language, and 126M shared ones.

The multilingual models are trained using the same hyperparameters as the monolingual ones. In this scenario, each training batch comprises data belonging to only one language. For the last three modifications, prior knowledge of the language in each recording is required during training and decoding.

4. Experimental work

4.1. Free data sources and overview of test sets

At present, two large Swedish datasets are freely available. The largest and most widely used is named NST. It was created in the early 2000s and targeted the development of dictation programs for the three Nordic languages. Later, the data was adopted by the Norwegian National Library and made publicly accessible. Its Swedish part[1] contains 480 hours of speech

[1] https://nb.no/sprakbanken/en/resource-catalogue/oai-nb-no-sbr-16

Table 1: *Swedish test set collection.*

test set	hours	words	words not seen [%]
NST5h	5.0	26,944	1.4
CV	6.3	36,922	0.5
SVT	4.1	37,056	1.9
PAR	3.0	25,884	0.7
YTB	0.6	5,601	3.7
ABOOK	1.0	6,904	8.0
FLEURS	2.3	15,507	5.4
all	22.3	154,818	2.0

Table 2: *Overview of Swedish training data (1,226 hours).*

train set	hours	style	envir.	availability
NST	402	read	clean	free
CV	41	read	mixed	free
PAR	349	mixed	mixed	harvested
SVT/YTB	73	mixed	mixed	harvested
ABOOK	361	read	clean	paid

(from almost 1,000 speakers). The set is large, but has some severe limitations. It is made of read sentences, out of which many are repeated by all of the speakers, and others contain just short commands or phonetically balanced (but not realistic) utterances. The data is officially split into the training and testing parts. From the former, we have sorted 402 hours for training after removing most of the repeated utterances. For testing, we randomly selected five hours (out of 100 hours available) to make this set's size comparable to the other ones. (Both the five-hour and full sets yield similar results.)

The second dataset is the Swedish part of Common Voice[2] (CV, version 9), which offers 41 hours for training and six hours for testing. It is made of read utterances, too, but with much-varying speaking styles and recording conditions.

As we want to develop and evaluate ASR systems aimed at various application domains, we have prepared several other test sets. The SVT set is made of 10 news shows broadcast by the eponymous TV channel in 2022. The shows are complete with one exception; non-Swedish spoken parts have been removed. The PAR set is made of talks that occurred in the parliament during the so-called interpellations in 2022. This set consists of 177 audio files, each spoken by a different person. The YTB set is a collection of recordings from several YouTube channels covering various topics. Another source is a short documentary audiobook dealing with international crime that contains many non-Swedish names and technical terms. It is narrated by a person not occurring in training. As the last item, we have included the test part of the recently popular FLEURS dataset [38].

The test sets altogether cover several speaking styles (read, planned, spontaneous, emotional), recording conditions (studio, large hall, home equipment), and various topics. Each file of these sets was transcribed automatically and then edited by a native speaker. Table 1 gives more information about the data, including the rates of the words not seen in the final training set.

4.2. Mono- and multi-lingual models trained on free data

We started our research by training an E2E model on the freely available Swedish data (NST+CV, 443 hours). As shown in the first row of Table 3, the WER values are reasonably low for the matched test sets (NST5h, CV) but, obviously, much higher for the two selected independent sets (SVT, PAR). Before launching a time-consuming process of harvesting additional training data, we implemented and evaluated all the modifications presented in Sec. 3. We combined the same Swedish data with data and models developed previously in our lab. We utilized an English model (trained on 10,000 hours of speech) and a Norwegian one (based on 900 hours collected from public sources in 2022 [39]). As expected, the models using Norwegian achieve the best results. It is mainly because the two languages (SWE and NO) are closely related; another reason is that, similar types of data sources were used. The exact impact of including Norwegian in the model-building process is presented in the rest of Table 3. All five investigated techniques yield significant WER reduction over the monolingual SWE model and are utilized in the following steps.

4.3. Automated data-harvesting process

In order to improve the E2E model, we have identified several Swedish public sources with large amounts of audio and associated texts. They, in particular, come from the Swedish radio and television company SVT with several channels and programs complemented by subtitles (mainly TV news) or text summaries (radio shows). Another relevant source is the Swedish parliament's archive of videos from plenary sessions and their official transcripts. There are also several Swedish YouTube channels with subtitles, although their number is still small. In general, the texts associated with these speech records are only loosely related, so we need a safe and reliable data harvesting scheme.

To broaden the scope of the training corpus, we have chosen 30 pairs of audio- and e-books (with respect to their topic, genre, narrator, and size) and purchased them in a downloadable format. The same harvesting approach is used here because the written and spoken text is not guaranteed to match exactly.

In our harvesting scheme, the data are processed as follows: First, the original (usually long) audio files are split into chunks shorter than 25 seconds. This is performed by the available ASR system, which also employs a voice activity detector [40] to determine suitable split points. The ASR output is aligned (using the Levenshtein distance method) with the provided reference text, which is then split into fragments assigned to the chunks. The second phase runs in iterations. In each, we use ASR system with all the model variants to transcribe the chunks and compare their outputs to the reference fragments. Those with a character error rate below 2% are added to the train set. Using multiple models increases the chance that more chunks meet the criterion. After processing all the available chunks, new variant

Table 3: *WER [%] of models (mono- and bi-lingual SWE/NO) based on freely available Swedish training data (NST+CV).*

model	NST5h	CV	SVT	PAR
mono SWE	5.9	7.5	30.2	24.1
init SWE from NO	4.6	**6.0**	**25.3**	**19.7**
joint SWE+NO	4.5	6.9	25.9	20.2
joint-with-lid SWE+NO	**4.4**	6.8	25.5	20.7
multi-decoder SWE+NO	4.5	6.5	25.7	19.9
multi-layer SWE+NO	4.5	6.2	26.2	20.3

[2] https://commonvoice.mozilla.org

Table 4: *WER [%] of SWE models (using NO as a support language) trained with increasing amounts of data across all test datasets.*

model	**25 h**	**50 h**	**75 h**	**100 h**	**150 h**	**200 h**	**250 h**	**300 h**	**500 h**	**1000 h**
mono SWE	66.4	39.8	28.1	22.8	18.1	16.1	14.7	13.8	12.0	10.8
init SWE from NO	44.7	28.9	23.2	19.2	14.9	**13.5**	**12.4**	12.3	**11.1**	**9.9**
joint SWE+NO	22.6	19.5	17.6	16.2	14.6	13.7	13.9	12.5	11.9	10.4
joint-with-lid SWE+NO	**22.1**	20.3	17.8	**16.0**	14.3	13.7	13.1	12.6	11.7	10.4
multi-decoder SWE+NO	28.4	21.1	17.8	17.0	**14.0**	**13.5**	12.7	**12.2**	11.6	10.3
multi-layer SWE+NO	24.0	**19.4**	**17.4**	17.1	14.5	**13.5**	12.9	**12.2**	11.8	10.1

models are trained. This procedure is repeated until the number of newly acquired training data drops below a reasonable level.

We applied this iterative scheme to all of the downloaded audio files. Within 10 iterations, we have collected 1,226 hours of diverse training data, whose structure is presented in Table 2. The process can be fully automated. However, it is rational to introduce minor human assistance into it to increase the amount of harvested data. In each iteration step, we have identified the most frequent errors (typically, foreign names in news and audiobooks) and reviewed at least a few corresponding files.

4.4. Effect of increasing amounts of data

To fully investigate which variant of the architecture (and when) is more suitable for bootstrapping a new language, we have randomly sampled the gathered data and trained all the variants with different amounts of data ranging from 25 to 1,000 hours. Each addition of data (e.g., from 25 to 50 hours) re-uses the previous ones (25 hours) to simulate better bootstrapping.

The results are presented in Table 4 in the form of weighted WER values across all of the test datasets (Table 1). They show a clear trend: multilingual training (in any of its modifications) helps tremendously in the beginning when the amount of target data is severely limited, and the data from the closely related language helps to fill in the gaps (e.g., the difference in WER values between the mono- and any multi-lingual model is over 35% when using only 25 hours of the target language). Around the mark of 200 hours, the performance starts to equalize, and initialization from a related language (*init*) becomes a cheaper option (less training time) and, in the end, even a slightly better one (lower WER value). In this case, the related language in multilingual training starts to cause more confusion. Out of the multilingual modifications, the proposed *multi-layer* one performs the best as it separates only the language-dependent parts of the model; but all of the modifications are beneficial. Finally, not even 1,000 hours is sufficient for a solely monolingual model to outperform any use of a related language.

4.5. Final model and comparison to commercial ASR

Eventually, we have trained the final model (on NVIDIA A40 GPU in 7 days) on the complete training corpus (1,226 hours) using the technique identified as the best-performing in Sec. 4.4, i.e., initialization by Norwegian model. The results are given in Table 5. Unsurprisingly, the lowest WER values appear for the read speech sets (NST5h, CV, and ABOOK), though the content of the last one is quite challenging, with 8% of unseen words. For the more realistic sets (SVT and PAR), we can observe significant improvements compared to the initial results in Table 3 (e.g., WER of the SVT set improved from 25.3% to 12.6%), given by the enlargement of the training set by the corresponding type of data. Moreover, the WER of the FLEURS set proves that the model can generalize well for data not seen in training.

Table 5: *WER [%] comparison to commercial solutions.*

test set	**our final**	**MS Azure**	**Google Cloud**
NST5h	**2.9**	5.5	24.9
CV	**5.9**	10.5	22.4
SVT	12.6	**10.8**	35.0
PAR	**7.3**	11.5	26.5
YTB	11.3	**10.1**	31.4
ABOOK	**3.9**	11.4	23.7
FLEURS	**12.4**	12.9	21.1
all	**8.0**	10.1	26.8

To evaluate our results in a broader context, we have submitted the test data to the two commercial services that have Swedish in their portfolio: Microsoft Azure (Speech to Text v. 02/2023) and Google Cloud (Speech-to-Text v. 02/2023). The achieved WER values are added to Table 5. As we have no information about the two systems and their resources, let us just briefly comment that our results slightly outperform those of Microsoft, while the Google ones are significantly worse.

5. Conclusions

In this work, we have developed a hybrid CTC/AED-based E2E ASR model for Swedish. At first, we have combined data from freely available datasets with several existing and two new multilingual techniques to iteratively harvest more than 1,200 hours of public training data. A closely related language – Norwegian – has been used as the supporting one. During this process, we have shown that multilingual training is more beneficial with limited data, while transfer learning becomes computationally cheaper and even further reduces the WER values when enough data is provided. After that, we have trained the final initialized model using the whole training set and the achieved results on average outperform two available commercial cloud services. The evaluation makes use of a diverse set of the testing data we have gathered. Moreover, we release the test sets or links to their sources on our cloud[3] along with detailed logs from the final tests, including ASR-to-reference word alignments.

6. Acknowledgements

The research has been funded by the EEA/Norway Grants and the Technology Agency of the Czech Republic (project No. TO01000027), and by the Student Grant Competition of TUL (project No. SGS-2022-3052). Computational resources were provided by the e-INFRA CZ project (ID:90140), supported by the Czech Ministry of Education, Youth and Sport.

[3] https://owncloud.cesnet.cz/index.php/s/MaoWzfXiRHp2DK0

7. References

[1] D. Wang, X. Wang, and S. Lv, "An overview of end-to-end automatic speech recognition," *Symmetry*, vol. 11, no. 8, 2019.

[2] V. Pratap, A. Sriram, P. Tomasello, A. Y. Hannun, V. Liptchinsky, G. Synnaeve, and R. Collobert, "Massively multilingual ASR: 50 languages, 1 model, 1 billion parameters," in *Interspeech 2020, Shanghai, China*, 2020.

[3] B. Li, R. Pang, Y. Zhang, T. N. Sainath, T. Strohman, P. Haghani, Y. Zhu, B. Farris, N. Gaur, and M. Prasad, "Massively multilingual ASR: A lifelong learning solution," in *ICASSP 2022, Singapore*, 2022.

[4] Y. Zhu, P. Haghani, A. Tripathi, B. Ramabhadran, B. Farris, H. Xu, H. Lu, H. Sak, I. Leal, N. Gaur, P. Moreno, and Q. Zhang, "Multilingual speech recognition with self-attention structured parameterization," in *Interspeech 2020, Shanghai, China*, 2020.

[5] M. Malmsten, C. Haffenden, and L. Borjeson, "Hearing voices at the national library - a speech corpus and acoustic model for the Swedish language," in *Fonetik 2022, Stockholm, Sweden*, 2022.

[6] R. Al-Ghezi, Y. Getman, A. Rouhe, R. Hilden, and M. Kurimo, "Self-supervised end-to-end ASR for low resource L2 Swedish," in *Interspeech 2021, Brno, Czechia*, 2021.

[7] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *NeurIPS 2020*, 2020.

[8] D. Wang and T. F. Zheng, "Transfer learning for speech and language processing," in *APSIPA 2015, Hong Kong*, 2015.

[9] J. Kunze, L. Kirsch, I. Kurenkov, A. Krug, J. Johannsmeier, and S. Stober, "Transfer learning for speech recognition on a budget," in *Rep4NLP@ACL 2017, Vancouver, Canada*, 2017.

[10] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *ICASSP 2013, Vancouver, Canada*, 2013.

[11] G. Heigold, V. Vanhoucke, A. W. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *ICASSP 2013, Vancouver, Canada*, 2013.

[12] W. Hou, Y. Dong, B. Zhuang, L. Yang, J. Shi, and T. Shinozaki, "Large-scale end-to-end multilingual speech recognition and language identification with multi-task learning," in *Interspeech 2020, Shanghai, China*, 2020.

[13] L. Chen and V. Leutnant, "Acoustic model bootstrapping using semi-supervised learning," in *Interspeech 2019, Graz, Austria*, 2019.

[14] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," in *Interspeech 2019, Graz, Austria*, 2019.

[15] A. Rosenberg, Y. Zhang, B. Ramabhadran, Y. Jia, P. J. Moreno, Y. Wu, and Z. Wu, "Speech recognition with augmented synthesized speech," in *ASRU 2019, Singapore*, 2019.

[16] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *ICML 2014, Beijing, China*, vol. 32, 2014.

[17] J. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *NIPS 2015, Montreal, Canada*, 2015.

[18] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiat, S. Watanabe, and T. Hori, "Multilingual sequence-to-sequence speech recognition: Architecture, transfer learning, and language modeling," in *SLT 2018, Athens, Greece*, 2018.

[19] A. Graves, "Sequence transduction with recurrent neural networks," in *ICML 2012*, 2012.

[20] V. Joshi, R. Zhao, R. R. Mehta, K. Kumar, and J. Li, "Transfer learning approaches for streaming end-to-end speech recognition system," in *Interspeech 2020, Shanghai, China*, 2020.

[21] M. Giollo, D. Gunceler, Y. Liu, and D. Willett, "Bootstrap an end-to-end ASR system by multilingual training, transfer learning, text-to-text mapping and synthetic audio," in *Interspeech 2021, Brno, Czechia*, 2021.

[22] L. Zhou, J. Li, E. Sun, and S. Liu, "A configurable multilingual model is all you need to recognize all languages," in *ICASSP 2022, Singapore*, 2022.

[23] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," in *ICASSP 2018, Calgary, Canada*, 2018.

[24] S. T. Abate, M. Y. Tachbelie, and T. Schultz, "End-to-end multilingual automatic speech recognition for less-resourced languages: The case of four Ethiopian languages," in *ICASSP 2021, Toronto, Canada*, 2021.

[25] S. Toshniwal, T. N. Sainath, R. Weiss, B. Li, P. J. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *ICASSP 2018, Calgary, Canada*, 2018.

[26] A. Kannan, A. Datta, T. N. Sainath, E. Weinstein, B. Ramabhadran, Y. Wu, A. Bapna, Z. Chen, and S. Lee, "Large-scale multilingual speech recognition with a streaming end-to-end model," in *Interspeech 2019, Graz, Austria*, 2019.

[27] V. Shetty, N. J. M. S. Mary, and S. Umesh, "Improving the performance of transformer based low resource speech recognition for Indian languages," in *ICASSP 2020, Barcelona, Spain*, 2020.

[28] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *ASRU 2017, Okinawa, Japan*, 2017.

[29] G. I. Winata, G. Wang, C. Xiong, and S. C. H. Hoi, "Adapt-and-adjust: Overcoming the long-tail problem of multilingual speech recognition," in *Interspeech 2021, Brno, Czechia*, 2021.

[30] Y.-A. Chung, Y. Zhang, W. Han, C.-C. Chiu, J. Qin, R. Pang, and Y. Wu, "w2v-bert: Combining contrastive learning and masked language modeling for self-supervised speech pre-training," in *ASRU 2021, Cartagena, Colombia*, 2021.

[31] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, "Unsupervised cross-lingual representation learning for speech recognition," in *Interspeech 2021, Brno, Czechia*, 2021.

[32] A. Babu, C. Wang, A. Tjandra, K. Lakhotia, Q. Xu, N. Goyal, K. Singh, P. von Platen, Y. Saraf, J. Pino, A. Baevski, A. Conneau, and M. Auli, "XLS-R: self-supervised cross-lingual speech representation learning at scale," in *Interspeech 2022, Incheon, Korea*, 2022.

[33] J. Bai, B. Li, Y. Zhang, A. Bapna, N. Siddhartha, K. C. Sim, and T. N. Sainath, "Joint unsupervised and supervised training for multilingual ASR," in *ICASSP 2022, Singapore*, 2022.

[34] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. E. Y. Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Interspeech 2018, Hyderabad, India*, 2018.

[35] A. Gulati, J. Qin, C.-C. Chiu, N. Parmar, Y. Zhang, J. Yu, W. Han, S. Wang, Z. Zhang, Y. Wu, and R. Pang, "Conformer: Convolution-augmented transformer for speech recognition," in *Interspeech 2020, Shanghai, China*, 2020.

[36] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," in *ACL 2016, Berlin, Germany*, 2016.

[37] T. Kudo and J. Richardson, "Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing," in *EMNLP 2018, Brussels, Belgium*, 2018.

[38] A. Conneau, M. Ma, S. Khanuja, Y. Zhang, V. Axelrod, S. Dalmia, J. Riesa, C. Rivera, and A. Bapna, "FLEURS: few-shot learning evaluation of universal representations of speech," in *SLT 2022, Doha, Qatar*, 2022.

[39] J. Nouza, L. Mateju, P. Cerva, and J. Zdansky, "Developing state-of-the-art end-to-end ASR for Norwegian," in *TSD 2023, Plzen, Czechia*, 2023.

[40] L. Mateju, F. Kynych, P. Cerva, J. Zdansky, and J. Malek, "Using x-vectors for speech activity detection in broadcast streams," in *Interspeech 2021, Brno, Czechia*, 2021.