INTERSPEECH 2023
20-24 August 2023, Dublin, Ireland

wav2vec 2.0 ASR for Cantonese-Speaking Older Adults in a Clinical Setting

Ranzo C. F. Huang, Brian Mak

Department of Computer Science and Engineering,
The Hong Kong University of Science and Technology, Hong Kong
{cfrhuang,mak}@cse.ust.hk

Abstract

The lack of large-scale speech corpora for Cantonese and older adults has impeded the academia's research of automatic speech recognition (ASR) systems for the two. On the other hand, the recent success of self-supervised speech representation learning has shown its competitiveness in low-resource ASR. This work therefore studies the application of wav2vec 2.0 ASR using monolingual and cross-lingual pre-trained models on a developing speech corpus, CU-MARVEL, which is dedicated to the automated screening of neurocognitive disorders (NCD) for Cantonese-speaking older adults in Hong Kong. We detail our data preparation procedures for creating a monolingual wav2vec 2.0 model from scratch and further pre-training a cross-lingual model. We report the performance of our wav2vec 2.0 ASR models on the said corpus and present a preliminary analysis of the relationship between the ASR performance of older adult speech and various demographic characteristics.

Index Terms: wav2vec 2.0, older adults, Cantonese, self-supervised speech representations, ASR

1. Introduction

Transformer-based networks pre-trained by self-supervised speech representation learning methodologies, such as wav2vec 2.0 [1], HuBERT [2], and data2vec [3], have learned to perform frame-level classification of pseudo-phonetic units automatically derived from unlabeled audio data. With an additional transfer learning stage coupled with some amount of labeled data, which is commonly referred to as 'fine-tuning', such models may be adapted to automatic speech recognition (ASR) tasks and achieve a performance superior to its supervised learning-only counterpart, especially in scenarios when a limited amount of labeled data are available [1, 2, 3]. The analysis of [4] on English wav2vec 2.0 models suggests pre-training lets the models learn to encode acoustic and linguistic information following an auto-encoder-like hierarchy; whereas fine-tuning with labels allows the models to encode phonetic and word information better at the highest few layers and breaks the structure's symmetry. The finding may suggest that while the information learned from a larger dataset during pre-training is exploitable by and hence favorable to low-resource ASR scenarios, the performance of the downstream ASR task still benefits from an increasing amount of (pseudo-)labeled data, as shown by, for example, [1, 5].

A practical use case that perfectly fits into the pre-training-fine-tuning framework is to provide automatic transcripts as an assistance to construct a large-scale conversational-style corpus, where audio recordings arrive much faster than transcriptions. Manual annotating this kind of speech data is both time-consuming and exhausting, particularly when precise and verbose transcripts are required for fine-grained linguistic analyses, e.g., Alzheimer's disease (AD) detection where filled pauses are utilizable features, e.g., see [6]. Self-supervised speech representation learning-based ASR eases the problem in part by learning from the unlabeled in-domain data in advance, and when more labeled data become available, the learned representations will help fine-tuning to produce better ASR results. Nevertheless, it takes time to collect enough data for starting off the training of self-supervised speech representations. Apart from that, for in-domain systems, capturing linguistic information from diverse data may be beneficial to cope with everyday speech.

The non-necessity of precise labels for self-supervised speech representation learning makes it easily extendable to a cross-lingual setting, where a speech representation model is trained with audio data in diverse languages, in the hope that the resultant representations are generalizable across multiple languages, thereby alleviating the need of language-specific models, which are expensive to build. XLS-R [7], an extension of XLSR [8], are a collection of cross-lingual wav2vec 2.0 models in three sizes (300M, 1B, and 2B parameters) pre-trained on 436K hours of speech data, of which the majority are in European languages. Cantonese and its variant Hong Kong Cantonese, which this work has a particular interest in, account for only 181 hours[1] of XLS-R's pre-training data and are therefore under-represented. It is then questionable if XLS-R's representations are well transferable to the language's ASR tasks. Moreover, vanilla cross-lingual speech representations may suffer from language interference. We therefore see cross-lingual representations as a prototyping tool for ASR fine-tuning while we develop monolingual pre-trained models in parallel.

In this work, we embrace both monolingual and cross-lingual speech representations, and study the ASR performance of the fine-tuned models stemming from (1) a Cantonese wav2vec 2.0 model pre-trained on out-domain data, (2) XLS-R, (3) the said Cantonese wav2vec 2.0 model further pre-trained on in-domain data, and (4) XLS-R further pre-trained on in-domain data, with the aim of saving the cost of in-domain development. The following sections will detail the data collection and preparation procedure, as well as the experimental setup and results, and will present our findings on the relationship between the demographic characteristics of older adults and the ASR performance of our best model.

[1] This figure considers the ISO language codes of *zh-HK* and *yue*. For their distinctions, please see the discussion on https://github.com/common-voice/common-voice/issues/2926 .

10.21437/Interspeech.2023-2470

2. Data Preparation

2.1. CU-MARVEL

CU-MARVEL [9] is an ongoing effort that targets a thousand Cantonese-speaking older adults (aged over 60 years) in Hong Kong and collects their spoken responses to a battery of NCD screening tests over time. After taking such assessments upon a visit, they are classified either as healthy, with mild NCD, or with major NCD.

This work considers only the data collected from the older adult participants' first visit, which we hereafter refer to as the 'baseline' data. The baseline data involves audio-recorded sessions in each of which an assessor conducts a list of NCD screening tests for an older adult participant in person in a room, which may be a sound-proof or a non-sound-proof one. On average, such a session is more than 1.5 hours long, and a participant speaks for around 30% of the time. It is worth noting that the manual transcriptions are done for a subset of the assessment tests only, and at the time of writing, the transcription work is still in progress.

We use both the labeled and unlabeled training data of the November 2022 release for system development, and evaluate the systems' performance on the labeled test data of the February 2023 release.[2] The breakdown of these data is provided in Table 1. Among the participants whom this work considers, the majority of them are aged below 80 years, and the number of female participants is 25% more than that of male participants. The age distribution of these participants is given in Figure 1.

Table 1: *Breakdown of the CU-MARVEL 'baseline' data.*

Split	**No. of sessions**	**Manually labeled hours**	
		Assessors	*Participants*
Partially labeled sessions			
Train (Nov 2022 ver.)	124	24.3	29.3
Test (Feb 2023 ver.)	46	13.0	14.8
Unlabeled sessions			
Train (Nov 2022 ver.)	288	-	-

2.2. Obtaining speech segments for wav2vec pre-training

We see diarization as a necessary pre-processing step for dividing unlabeled audio data into utterances, so that the pre-training data better matches the ASR data. Below we describe our procedures for preparing the in- and out-domain pre-training data.

2.2.1. *Cantonese older adult speech data*

With a large number of unlabeled training sessions available, together with the unannotated regions of the partially labeled training sessions, we consider them a rich source of data for wav2vec 2.0 pre-training. To obtain diarized speech segments from the unlabeled data, we further train the pre-trained segmentation pipeline[3] from pyannote.audio [10, 11] with supervisions derived from the labeled data annotations. The end-to-end SincNet-LSTM-based segmentation model is further trained on 5-second chunks with speech activities from at most 2 speakers. Further training the model for 10 epochs using 4 NVIDIA Quadro RTX8000 GPUs on a private server took an hour. We repeated the same method on another Cantonese older adult speech corpus, CUHK-JCCOCC-MoCA [12], to obtain slightly more data on top of CU-MARVEL's baseline data. After combining the labeled data and the automatically segmented data from the two corpora, we obtained 503 hours of speech segments for wav2vec 2.0 pre-training.

2.2.2. *Cantonese out-domain speech data*

Since there does not exist any pre-defined data in Cantonese for self-supervised speech representation learning, we seek to create our own set of data by obtaining audio data from the web. To gain better control of audio quality, we pool data from a limited number of sources, namely 59 Cantonese podcast shows and 1 YouTube channel. The sources provide mostly conversational-based content, including casual chats, interviews, discussions, and Skype call-ins. To the best of our knowledge, most of the speakers are not older adults. Most podcast recordings are in the MP3 format with a sampling rate of 44.1 kHz, and the audio recordings downloaded from YouTube are in the Opus format with a sampling rate of 48 kHz. The data are resampled to 16 kHz and stored in the FLAC format.

Without a diarization training dataset that matches the data's domain and language, we resort to simulating conversations using all data from the *zh-HK* and *yue* language subsets of Common Voice 11.0[4] excluding the data from the test speakers. We simulate conversations by randomly generating turn-hold, turn-switch, interruption, and backchannel transitions using the algorithm as described in [13]. Prior to creating simulated mixtures, we train a Kaldi [14] SAT GMM-HMM system to clean the Common Voice data and obtain word-level alignment information to determine the time boundaries of speech activities. The alignment information is also used in chunking the utterances into smaller fragments which are to serve as backchannel speech. A total of 884K mixtures which amount to 4.2K hours were created to simulate 1- to 4-speaker mixtures. To deal with the actual conversations with a variable number of speakers, we adopt the self-attention-based SA-EEND with encoder-decoder-based attractors [15] and icefall's[5] implemen-

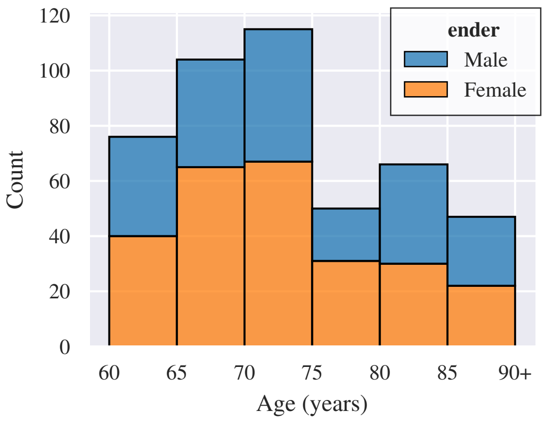

Figure 1: *Age distribution of the older adult participants of CU-MARVEL considered in this work.*

[2] The difference between the two releases is due to the amount of labeled data. There is no duplicate of participants in the training and test data.

[3] https://huggingface.co/pyannote/segmentation

[4] https://commonvoice.mozilla.org/en/datasets

[5] https://github.com/k2-fsa/icefall

tation of 'Reworked' Conformer as the backbone. The input features are 80-dimensional log-mel filterbank coefficients with a frame shift of 10 ms and a frame length of 25 ms. The encoder consists of a 3-layer 2D-CNN sub-sampling module, which subsamples the input sequence by a factor of 10, 4 Conformer layers with a dimension of 384 and 6 heads, followed by 1 layer of uni-directional LSTM accepting chronologically shuffled encoded features; the decoder is a 1-layer uni-directional LSTM. The model has 20.7M parameters. Training the model for 20 epochs using 2 NVIDIA RTX A6000 GPUs on a rented server took 1 day.

Segmentation of the raw data is done with SpeechBrain's[6] neural VAD[7] front-end and the said SA-EEND model. We kept only the segments that are at least 2 seconds and at most 40 seconds long. The segmentation procedure at the end produced 2.8K hours of speech segments for wav2vec 2.0 pre-training.

3. Experiments

3.1. wav2vec 2.0 pre-training

3.1.1. Cantonese wav2vec 2.0 pre-training

This experiment aims to produce a monolingual pre-trained model for comparing the downstream ASR performance yielded from monolingual and cross-lingual speech representations. We use fairseq's [16] implementation of CNN-Conformer for the pre-trained model. We use 12 Conformer layers, each with a dimension of 768 and 12 attention heads. This gives rise to 180M learnable parameters. Here, we use the data as described in Section 2.2.2. We train the model for 320K steps, or 96 epochs, using FP16 training and the AdamW optimizer with a weight decay of 0.01. We set the learning rate to 3e-4, and warm up for 10% of the training steps, and follow a linear decay schedule. The mask probability is set to 0.65 and the mask length is set to 10. Pre-training using 6x NVIDIA RTX A6000 GPUs on a rented server took 8 days.

3.1.2. Cantonese wav2vec 2.0 further pre-training

This experiment performs domain adaptation on a monolingual speech representation model. We use fairseq to further pre-train the Cantonese wav2vec 2.0 model (Section 3.1.1) on the data described in Section 2.2.1. We freeze the CNN layers, and train only the Transformer model and the quantization modules for 80K steps, which amounts to 96 epochs, using FP16 training and the AdamW optimizer with a weight decay of 0.01. We use a learning rate of 2e-4 with no warm-up and follow a linear decay schedule. The masking configuration is the same as in Section 3.1.1. This took us 5 days to complete the pre-training using 3x NVIDIA Quadro RTX 8000 GPUs on a private server.

3.1.3. XLS-R further pre-training

This experiment helps the study of the use of cross-lingual speech representations for the fast prototyping of an in-domain speech representation model. We use fairseq to further pre-train the 300M XLS-R on the data as described in Section 2.2.1. The model is a CNN-Transformer that possesses 24 Transformer layers, each with a dimension of 1024 and 16 attention heads. The training configuration follows Section 3.1.2. This took us 7 days to complete the pre-training using 3x NVIDIA Quadro RTX 8000 GPUs on a private server.

3.2. wav2vec 2.0 ASR fine-tuning

Using the labeled data of CU-MARVEL, we obtain ASR models by fine-tuning (1) the Cantonese Conformer wav2vec2.0 Section 3.1.1), (2) the 300M XLS-R, (3) the further-pre-trained Cantonese wav2vec 2.0 (Section 3.1.2), and (4) the further-pre–trained XLS-R (Section 3.1.3) for recognition performance comparison.

We consider building phone lexicon-based ASR systems due to the size of the available labeled data and the large character space of Cantonese Chinese, and base our lexicon on the pronunciation dictionary from words.hk[8] and Jyutping Table[9]. A problem with adopting a phone-based lexicon is that some words come with multiple pronunciations, and we do not know which one of them a speaker is actually referring to. Moreover, imperfect word segmentation for the training transcripts adds further ambiguities. Therefore, we use k2's[10] implementation of CTC which supports graph-based supervisions that allow the inclusion of alternative pronunciations of words for computing the CTC loss.

We apply the following fine-tuning configuration unanimously to the three pre-trained models: we freeze the CNN module and fine-tune the other parts of the model for 40K steps, or 190 epochs, using FP16 training and the AdamW optimizer without weight decay; the learning rate is set to 3e-5, with a tri-stage schedule as adopted by [1], in which the first 10% of training steps are for warm-up and training the output layer only, the next 40% are for a constant learning rate, and the remaining steps are for linearly decaying the learning rate; we use a mask probability of 0.75, and a layer-drop probability of 0.1. We trained each model using 2 NVIDIA RTX Quadro RTX 8000 GPUs on a private server, and training each took half a day.

4. Results and Analysis

The output of the fine-tuned models is decoded using word n-gram LMs with the method of whole-lattice re-scoring from icefall. The performances of the three models are given in Table 2, with a detailed breakdown by gender (*male* or *female*), speaker role (*assessors* or *participants*), and recording environment (*sound-proof* or *non-sound-proof*). Below we compare their performance and analyze the implications.

4.1. Monolingual vs. cross-lingual representations

Comparing the performances of **Model 1** (Cantonese Conformer) and **Model 2** (300M XLS-R), we observe that the monolingual model offers some but consistent improvement upon the cross-lingual model in all aspects. This suggests that ASR fine-tuning benefits from a pre-trained model which matches the language of the fine-tuning data. The improvement, however, is limited by other factors. Environmental robustness of the model is an issue, at least because the improvement of recognizing assessor speech in a non-sound-proof venue is only half of that in a sound-proof venue (a reduction of 6.06% vs. 10.95%). Another problem may be attributed to the exclusion of older adult data during pre-training, as seen from the fact that the assessors enjoyed a greater reduction in overall CER than the older adult participants (a reduction of 7.92% vs. 5.30%).

[6] https://github.com/speechbrain/speechbrain

[7] https://huggingface.co/speechbrain/vad-crdnn-libriparty

[8] https://words.hk/faiman/analysis

[9] https://github.com/lshk-org/jyutping-table

[10] https://github.com/k2-fsa/k2

Table 2: *Character error rate (%) on the baseline test data of CU-MARVEL (Feb 2023 ver.)*

Model	**Sound-proof venue**				**Non-sound-proof venue**				**Overall**	
	Assessors	*Participants*			*Assessors*	*Participants*			*Assessors*	*Participants*
	Overall	Male	Female	Overall	Overall	Male	Female	Overall		
(1) Cantonese Conformer	4.97	17.57	14.65	15.52	7.06	26.78	17.77	22.45	6.11	18.91
(2) XLS-R (300M)	5.59	18.88	16.50	17.21	7.51	26.94	18.42	22.85	6.64	19.96
(3) (1) + further pre-train	3.98	17.11	14.33	15.17	5.33	25.30	16.32	20.99	4.72	18.01
(4) (2) + further pre-train	**3.88**	**15.95**	**13.59**	**14.29**	**4.97**	**21.84**	**14.56**	**18.34**	**4.47**	**16.27**

4.2. Effects of further pre-training

Comparing **Model 1** (Cantonese Conformer) and **Model 3** (Cantonese Conformer, further pre-trained on in-domain data), as well as **Model 2** (300M XLS-R) and **Model 4** (300M XLS-R, further pre-trained on in-domain data), the further pre-trained models significantly improve upon the unadapted models, suggesting the crucial importance of including in-domain data during pre-training. The further pre-trained XLS-R outperforms the further pre-trained monolingual model, and we hypothesize that it is due to the larger model size of the XLS-R.

Consider the XLS-R pair, with further pre-training the assessor speech shows an overall 32.61% improvement, whereas the participants' shows 18.50%. One reason why the accuracy of recognizing the assessor speech greatly improves is that the speakers are seen during training (but not the participants), and they use consistent wordings throughout different assessment sessions to give instructions to the participants. We also witness a significant reduction of CER when recognizing the participant speech in a non-sound-proof venue (19.73%), which is much more prominent than that for a sound-proof venue (16.94%). This suggests the simple method of further pre-training allows the model to gain environmental robustness without the need of sophisticated tricks. However, there still exists a large performance gap in recognizing speech in a non-sound-proof venue when compared to a sound-proof venue: the CER for the former environment is more than 20% higher than the latter.

5. Discussion

To understand the relationship between the ASR performance on the participant speech and their demographics, as well as the recording environment, we fit a linear regression model to predict the participants' CER obtained with our best model (**Model 4**), from their gender, NCD classification, age, education years, and the recording condition. We set a significance level of $p < 0.02$. An F-test of the least-squares fit shows the overall regression model is significant ($p = 0.000$). Although age shows a weak positive correlation with CER ($r = 0.366$), there is a lack of support that an increasing age gives rise to higher CER ($p = 0.543$). On the other hand, while receiving more education years shows a very weak negative correlation ($r = -0.273$) with CER, education is not a significant factor to explain CER ($p = 0.028$). These findings suggest other factors are responsible for the correlations. Indeed, the F-test shows the other factors are significant variables ($p < 0.02$): the participant being a man ($p = 0.000$), the participant having a higher NCD severity level ($p = 0.017$), and the recording environment being a non-sound-proof one ($p = 0.012$) show a positive relationship with CER. These preliminary results suggest that NCD speech is more difficult to recognize. However, due to a small sample size, a follow-up investigation is needed to further verify the claim. In addition, male participant speech is also more difficult to recognize, possibly due to the imbalance of the participants' genders, suggesting the need of transcribing more male participant sessions to balance the data. Finally, environmental robustness is still an issue to resolve.

6. Conclusions

Specifically targeting an ASR application in a clinical setting, this work presents an example of a full pipeline for self-supervised representation learning with wav2vec 2.0, which includes diarization, (further) wav2vec 2.0 pre-training, and ASR fine-tuning. Our findings suggest the nature of the pre-training data is crucial to achieving a good performance in the downstream ASR system, and the determinants include, but not limited to language, speakers' age group, and recording environment. In the future, we may look for more sophisticated methods to incorporate more older adult data for pre-training and ease the problem of unbalanced demographics in the in-domain dataset.

7. Acknowledgements

The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. T45-407/19N-1).

8. References

[1] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., vol. 33. Curran Associates, Inc., 2020, pp. 12 449–12 460.

[2] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 29, pp. 3451–3460, 2021.

[3] A. Baevski, W.-N. Hsu, Q. Xu, A. Babu, J. Gu, and M. Auli, "data2vec: A general framework for self-supervised learning in speech, vision and language," in *Proceedings of the 39th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvari, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, Jul. 2022, pp. 1298–1312.

[4] 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.

[5] Q. Xu, A. Baevski, T. Likhomanenko, P. Tomasello, A. Conneau,

R. Collobert, G. Synnaeve, and M. Auli, “Self-training and pre-training are complementary for speech recognition,” in *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021, pp. 3030–3034.

[6] A. Balagopalan, B. Eyre, J. Robin, F. Rudzicz, and J. Novikova, “Comparing pre-trained and feature-based models for prediction of Alzheimer’s disease based on speech,” *Frontiers in Aging Neuroscience*, vol. 13, 2021.

[7] 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 *Proc. Interspeech 2022*, 2022, pp. 2278–2282.

[8] A. Conneau, A. Baevski, R. Collobert, A. Mohamed, and M. Auli, “Unsupervised cross-lingual representation learning for speech recognition,” in *Proc. Interspeech 2021*, 2021, pp. 2426–2430.

[9] H. Meng, B. Mak, M.-W. Mak, H. Fung, X. Gong, T. Kwok, X. Liu, V. C. T. Mok, P. Wong, J. Woo, X. Wu, K. H. Wong, S. S. Xu, N. Zheng, R. C. F. Huang, J. Kang, X. Ke, J. Li, J. Li, and Y. Wang, “Integrated and enhanced pipeline system to support spoken language analytics for screening neurocognitive disorders,” in *Interspeech 2023*, 2023.

[10] H. Bredin, R. Yin, J. M. Coria, G. Gelly, P. Korshunov, M. Lavechin, D. Fustes, H. Titeux, W. Bouaziz, and M.-P. Gill, “pyannote.audio: Neural building blocks for speaker diarization,” in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 7124–7128.

[11] H. Bredin and A. Laurent, “End-to-end speaker segmentation for overlap-aware resegmentation,” in *Proc. Interspeech 2021*, 2021, pp. 3111–3115.

[12] S. S. Xu, M.-W. Mak, K. H. Wong, H. Meng, and T. C. Kwok, “Speaker turn aware similarity scoring for diarization of speech-based cognitive assessments,” in *2021 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2021, pp. 1299–1304.

[13] N. Yamashita, S. Horiguchi, and T. Homma, “Improving the naturalness of simulated conversations for end-to-end neural diarization,” in *Proc. The Speaker and Language Recognition Workshop (Odyssey 2022)*, 2022, pp. 133–140.

[14] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz, J. Silovsky, G. Stemmer, and K. Vesely, “The Kaldi speech recognition toolkit,” in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding*, Dec. 2011.

[15] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue, and K. Nagamatsu, “End-to-End Speaker Diarization for an Unknown Number of Speakers with Encoder-Decoder Based Attractors,” in *Proc. Interspeech 2020*, 2020, pp. 269–273.

[16] M. Ott, S. Edunov, A. Baevski, A. Fan, S. Gross, N. Ng, D. Grangier, and M. Auli, “fairseq: A fast, extensible toolkit for sequence modeling,” in *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, Jun. 2019, pp. 48–53.