Two Front-Ends, One Model : Fusing Heterogeneous Speech Features for Low Resource ASR with Multilingual Pre-Training

Anonymous ACL submission

Abstract

Transfer learning is widely applied in various deep learning-based speech tasks, especially for tasks with a limited amount of data. Recent studies in transfer learning mainly focused on either supervised or self-supervised perspectives. This work, however, seeks to incorporate the two schemes together towards low-resource automatic speech recognition (ASR) for minor and endangered language (EL) communities. We propose a general framework to use learned transformations to resolve time resolution differences between any speech features, allowing for fusion of any self-supervised representations or spectral features used in multilingual pre-training. Our experiments over two low-resource languages and three ELs demonstrate that the proposed framework can significantly improve the absolute average word error rate from 45.4% to 35.5%.

1 Introduction

End-to-end (E2E) approaches to ASR have shown promising results compared to hybrid approaches for not only high-resourced scenarios (Chiu et al., 2018; Karita et al., 2019; Pham et al., 2019; Guo et al., 2021), but also certain low-resource scenarios in which linguistic documentations are insufficient for building lexicon-dependent models (Grenoble et al., 2011; Zahrer et al., 2020; Shi et al., 2021a). On the other hand, end-to-end approaches to low-resource ASR are distinctly disadvantaged by a lower data efficiency (Lüscher et al., 2019) and language-mismatch with powerful self-supervised representations (Hsu et al., 2021).

One direction towards mitigating these low-resource issues is to incorporate knowledge from several languages into multilingual end-to-end models (Watanabe et al., 2017; Toshniwal et al., 2018; Kannan et al., 2019). When there is no training data available for the target languages, these systems can be applied in a zero-shot manner (Li et al., 2020; Yan et al., 2021; Xu et al., 2021). Fortunately, many languages have small amounts of data which can be used to fine-tune large-scale multilingual models towards target languages, resulting in further improvements (Hou et al., 2020; Pratap et al., 2020; Adams et al., 2019; Li et al., 2021).

Another direction is to use self-supervised learning representations (SSLR) trained on large untranscribed corpora as a front-end feature for ASR, replacing conventional spectral features like log Mel filterbank coefficients (FBank) (Yi et al., 2020; Wu et al., 2020; Baevski et al., 2020; N et al., 2021; Chang et al., 2021; Liu et al., 2021). Although these approaches have shown improvements across many languages, performance depends on the relatedness between the SSLR training languages and the target language (Conneau et al., 2019).

In this work, we are interested in leveraging both multilingual pre-trained (MPT) models with conventional speech feature front-ends and various SSLRs as resources for our low-resource ASR systems. In particular, we seek to efficiently incorporate multiple speech features, which may or may not have the same time resolution, as fused inputs to our end-to-end models. We propose a general framework to fuse such heterogeneous speech features and investigate several different learnable transformations for the fusion(Sec 3). Then we describe one instance following this front-end fusion framework which combines HuBERT features (Hsu et al., 2021) with an MPT model trained on FBank features (Sec 4.2). We demonstrate experimentally that our method improves absolute average WER by 9.9% on three endangered languages, and two of other low-resource languages in Sec 4.3.

Further, our data, pre-trained models, and reproducible methods are released open-source[1] to promote future developments on several endangered (Totonac, Yoloxóchitl Mixtec, and Highland Puebla Nahuatl) and low-resourced (Arabic and Tamil) lan-

[1] Available after the double-blind review period.

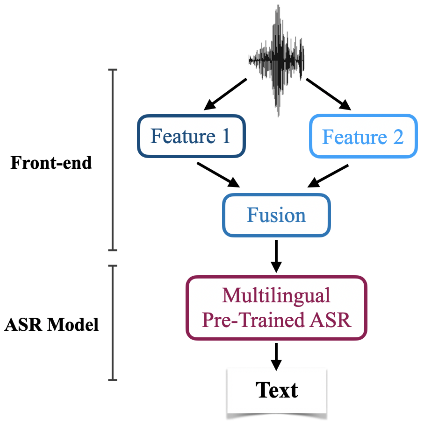

Figure 1: The architecture of our proposed model. The general formulation can be found in Sec 3.1.

guages. Notably, our released Totonac ASR data is the first publicly available annotated speech corpus.

2 Motivation

Multilingual models have not only reached large-scales, but they have also demonstrated high efficiency in modeling multiple languages. For instance, (Li et al., 2021) found multilingual training boosted low-resourced languages while also avoiding degradation of high-resource languages. Separately, SSLRs encode general-purpose information about speech that can apply to various downstream tasks, including ASR (Yang et al., 2021). For low-resource scenarios, Yi et al. (2020) found that wav2vec2 was useful for 6 low-resource languages, suggesting that SSLR can replace spectral features like FBanks in these cases.

Rather than viewing MPT and SSLR as two distinct techniques, we argue that for low-resourced ASR these ideas are critically intertwined for several reasons. (1) It is difficult to maintain broad compatibility of supervised MPT models if different front-ends are preferred for different scenarios. (2) For low-resource ASR, there are often domain mismatches between SSLR pre-training data and ASR fine-tuning for a target language, leading to unstable performances over when relying only on only an SSLR front-end. (3) SSLRs are often pre-trained exclusively over major languages like English (e.g., Hubert), again leading to unstable performances depending on cross-lingual similarities.

3 Methodology

3.1 Heterogeneous Speech Feature Fusion

Historically, multi-layer perceptron tandem features are concatenated with spectral features to reach better performances (Hermansky et al., 2000; Zhu et al., 2004; Lal and King, 2013). Following their insights, Chen et al. (2021) performed the concatenation between STFT and SSLR features for speech separation by repeating the SSLR feature across the time-domain. While this repeat-based method ameliorates the dimension mismatch of features, it does not necessarily produce optimally fused features for a particular task at hand.

We propose a general framework to fuse any SSLRs and spectral features through learnable transformations, allowing for the joint use of different supervised pre-training models with self-supervised representations. Such learnable fusions have been employed in various multi-source/multimodal applications previously (Libovický and Helcl, 2017; Hori et al., 2017). The framework is illustrated in Figure 1 and formulated as follows:

For an utterance, denote $\mathbf{X}^{\mathcal{F}_1} = (\mathbf{x}_t^{(1)} \in \mathbb{R}^{D_1} | t = 1, \cdots, T_1)$ as a feature type $\mathcal{F}_1$, where D_1 is the dimension of the feature at each frame and T_1 stands for the number of frames. Similarity, we could define $\mathbf{X}^{\mathcal{F}_2}$ along with D_2 and T_2. As D_1 and D_2, T_1 and T_2 may not necessarily be the same, combining two front-ends cannot be achieved by simply concatenation. In Figure 1, we introduce a fusion block to fuse features from two heterogeneous features (noted $\mathcal{F}_1$ and $\mathcal{F}_2$). Specifically, we perform learnable transformations over speech features $\mathbf{X}^{\mathcal{F}_1}$ and $\mathbf{X}^{\mathcal{F}_2}$ with linear, recurrent, convolution, or attention-based neural architectures. We then use RESHAPE so that the transforms of $\mathbf{X}^{\mathcal{F}_1}$ and $\mathbf{X}^{\mathcal{F}_2}$ have the same dimensions, as shown in Equation (1).

$$\begin{aligned} \tilde{\mathbf{X}}^{\mathcal{F}_1} &= \text{RESHAPE}(\text{TRANSFORM}(\mathbf{X}^{\mathcal{F}_1})) \\ \tilde{\mathbf{X}}^{\mathcal{F}_2} &= \text{RESHAPE}(\text{TRANSFORM}(\mathbf{X}^{\mathcal{F}_2})) \end{aligned} \quad (1)$$

After that, $\tilde{\mathbf{X}}^{\mathcal{F}_1}$ and $\tilde{\mathbf{X}}^{\mathcal{F}_2}$ are concatenated at the feature-dimension. In the next sub-section, we introduce one particular instance of this framework.

3.2 Fusion between SSLR and FBank for Multilingual Pre-trained (MPT) ASR

As discussed in Sec 2, several issues may raise when performing a simple combination between MPT and SSLR, though they can improve low-resource ASR, respectively. Therefore, we seek to fuse the requisite front-ends for our end-to-end model, following the formulation in Sec 3.1. Firstly, we pre-train a multilingual encoder-decoder model with language identification and hybrid CTC/Attention objectives (Watanabe et al., 2017; Hou et al., 2020). This architecture is built upon the FBanks, so we define our first front-end feature $\mathbf{X}^{\text{FB}} = (\mathbf{x}_t^{\text{FB}} \in \mathbb{R}^{D_{\text{FB}}} | t = 1, \cdots, T_{\text{FB}})$. Secondly, we use HuBERT as our SSLR front-end feature (Hsu et al., 2021), $\mathbf{X}^{\text{HUB}} = (\mathbf{x}_t^{\text{HUB}} \in \mathbb{R}^{D_{\text{HUB}}} | t = 1, \cdots, T_{\text{HUB}})$. Following our framework in Eq. (2), we compute $\tilde{\mathbf{X}}^{\text{FB}}$ and $\tilde{\mathbf{X}}^{\text{HUB}}$ before ultimately obtaining fused features $\mathbf{X}^{\text{FUSE}}$ as follows:

$$\mathbf{X}^{\text{FUSE}} = \tilde{\mathbf{X}}^{\text{FB}} \oplus \tilde{\mathbf{X}}^{\text{HUB}}, \tag{2}$$

where $\oplus$ denotes feature-dimension concatenation.

4 Experiments

4.1 Datasets

We use a combination of Commonvoice 5.1 (Ardila et al., 2020) and Voxforge (www.voxforge.org) for MPT. The corpus results in 5,029 hours of training data, including 52 languages from different language families.

We enroll three endangered languages which are not included in the multilingual corpus for testing, including Yoloxóchitl Mixtec (YM), Highland Puebla Nahuatl (HPN), and Totonac. Though endangered, YM and HPN have around 100 hours of transcribed speech in their released version (Shi et al., 2021b,a). However, to simulate a low-resource scenario, we randomly select 5,000 utterances (around 10 hours) from the official training sets, but used the same validation and test sets, as introduced in (Shi et al., 2021b,a). Totonac is another EL, spoken in the northern sierras of Puebla and adjacent areas of Veracruz. In this work, we release a public available version of Totonac speech resources. The corpus includes 10 hours of speech (86 long recordings) with fine-grained transcription. We randomly select 70 recordings as the training set, 8 for validation, and 8 for testing.

In addition to the three endangered languages mentioned above, we perform experiments on Arabic (AR) and Tamil (TA) corpora from Commonvoice to assert the robustness of our proposed methods in an in-domain low resource scenario. Both Arabic and Tamil have 20 hours of speech.

4.2 Experimental Setups

Baseline (A): For all the languages, our baseline (namely Exp **A** in later sections) adopts the same transformer-based encoder-decoder architecture with CTC/Attention hybrid training (Kim et al., 2017). The front-end in Exp **A** extracts FBank features at a frame length of 20ms and a frameshift of 8ms. The extracted FBank features are firstly subsampled with a convolutional block and then fed into the encoder-decoder. The encoder contains 12-layer self-attention blocks with 4-head attention and 512-dimensional hidden sizes. While the decoder has 6 cross-attention transformer blocks. Specaugmentation (Park et al., 2019) and speed perturbation are employed for data augmentation. For training, we use Adam optimizer and Noam scheduler with a 1.0 learning rate at peak. The warm-up step is set to 4,000, considering the low-resource scenario. All the parameters are initialized with Xavier uniform distribution (Glorot and Bengio, 2010). The ASR model is trained on byte-pair-encoding (BPE) units of 250. The same architecture and training configuration are aligned for the following experiments.

Multilingual Pre-training (B): MPT is performed on a large-scale corpus introduced in Sec 4.1. We follow the same pre-training strategy as (Hou et al., 2020). For each utterance, the model needs to generate a language ID token prior to ASR transcription. To keep a necessary coverage for transcribing all 52 languages, we set a BPE size of 7,000. Large batch size is applied here in order to stabilize the multilingual training. After the pre-training, Exp **B** is conducted with parameter initialization from the pre-trained model.

Self-supervised Representation (C): In our experiments, we employ HuBERT as the front-end.[2] To fully explore the potential of HuBERT, we select the HuBERT-large model pre-trained over 60k hours of LibriLight (Kahn et al., 2020; Ott et al., 2019). The SSLR wrapper provided in (Yang et al., 2021) is applied to extract high-dimensional features with 20ms frameshift. In Exp **C**, the model directly applies HuBERT representation for training, which is the same approach as in (Chang et al., 2021). As for ablation purposes, we also conduct experiments on only SSLRs front-end with the initialization from the MPT model. We name these experiments as Exp **C'** in the next section.

Joint-system (D&E): The joint-system incorporates both FBank and SSLR in model front-end. According to our settings, the resolution ratio be-

[2] We also conduct experiments on Wav2vec2 and Wav2vec2-XLSR (Conneau et al., 2019; Baevski et al., 2020). However, the performances are not stable for ASR training.

	Front-end			ASR	CER/WER					
Exp	FBank	SSLR	Align	MPT	Totonac	YM	HPN	AR	TA	Avg.
A	✓	✗	-	✗	17.3/50.6	26.2/50.8	51.5/77.6	15.4/29.2	6.1/**19.0**	23.3/45.4
B	✓	✗	-	✓	17.9/50.3	24.8/47.5	34.4/64.6	12.9/26.7	8.2/24.0	19.6/42.6
C	✗	✓	-	✗	17.1/48.3	38.8/61.2	29.4/58.3	15.1/29.2	6.2/19.7	21.3/43.3
D	✓	✓	✓	✗	14.6/46.7	**19.1**/42.6	**23.1**/52.4	**8.4**/22.4	6.1/24.5	**14.3**/37.7
E	✓	✓	✓	✓	**14.4/45.6**	20.0/**40.0**	25.1/**52.1**	9.2/**20.2**	**5.9**/19.4	14.9/**35.5**

Table 1: Results comparing our proposed fused front-end models (**D**, **E**) with various single front-end baselines (**A**, **B**, **C**) for 5 low-resourced or endangered languages, as measured by Character (CER) and Word (WER) Error Rates.

Exp	Front-End	MPT	CER	WER
A	FBANK	✗	17.3	50.6
B	FBANK	✓	17.9	50.3
Δ (**A**→**B**)	-	-	-0.6	+0.3
C	HUBERT	✗	17.1	48.3
C'	HUBERT	✓	20.9	58.4
Δ (**C**→**C'**)	-	-	+3.8	+10.1

Table 2: Ablation study comparing improvement/degradation on Totonac when incorporating FBank-based MPT with FBank-based fine-tuning ($\Delta(\mathbf{A} \rightarrow \mathbf{B})$) vs. incorporating FBank-based MPT with HuBERT-based fine-tuning ($\Delta(\mathbf{C} \rightarrow \mathbf{C'})$).

Exp	Fusion type	CER	WER
D	Linear	**14.6**	**46.7**
D1	Repeat	15.8	48.2
D2	RNN	65.1	86.9
D3	Convolution	16.1	50.8
D4	Attention	17.8	52.4

Table 3: Ablation study comparing performance on Totonac of several fusion types (**D1-4**) with our proposed linear fusion model without ASR pre-training (**D**).

tween FBank and HuBERT feature is 5:2. As discussed in Sec 3.1, we apply transformations to both features and then reshape them into the same time resolution. The linear layer contains 400 units in our experiments. The model then consumes the concatenation of both HuBERT and FBank features as inputs. We name the experiments with the fusion block as Exp **D**. We refer to the experiments as Exp **E** if it is initialized with the MPT model. We default to using the linear fusion block. But, to investigate other potential methods for feature fusion, we conduct ablation studies (i.e., Exp **D1**-**D4**) over four other approaches, including simple repeating, convolution, recurrent, and attention-based fusion.

4.3 Results and Discussion

Table 1 provides results of our main experiments over the five low resource languages introduced in Sec 4.1. Our proposed model (Exp **E**) reaches the best performances, which improves 8.4% absolute average CER and 9.9% absolute average WER than the baseline in Exp **A**. Besides, Exp **B** with MPT model leads to notable improvements over the baseline for some languages such as HPN, even though HPN was not in the set of languages used by the multilingual pre-training model. SSLR could also benefit some languages (e.g., Totonac) as indicated from Exp **C**. According to Exp **D**, the proposed fusion module demonstrates better performances for most languages, and also reaches the best average CER across the five languages.

Table 2 shows ablation study of Exp **C'** where we only use SSLR features with initialization from the MPT model originally built upon FBanks. Exp **C'** is degraded compared to Exp **A**, **B**, and **C**, suggesting incompatibility between SSLR and MPT model as the latter is trained on FBank features.

Table 3 provides results for Exp **D**, which consider the various fusion strategies discussed in Sec 3.1. It shows that linear fusion outperforms simple repeat method (i.e., Exp **D1**). Recurrent, convolution and attention-based networks strategies are also less effective than the linear approach in our context.

5 Conclusion

In this work, we suggest that self-supervised learning and supervised pre-training can jointly improve the ASR performances in low-resource scenarios. We propose a framework to align features with different time-domain resolutions and demonstrate the effectiveness of fusing various front-ends features. We also release a Totonac ASR corpus, serving for the purpose of endangered language documentation, and we show that our reproducible methods enable to get very good results in very low resource scenarios. In future works, we will investigate (1) multilingual pre-training with fused SSLR features; (2) zero-shot learning, especially for EL documentation purposes.

References

Oliver Adams, Matthew Wiesner, Shinji Watanabe, and David Yarowsky. 2019. Massively multilingual adversarial speech recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 96–108.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *LREC*.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33.

Xuankai Chang, Takashi Maekaku, Pengcheng Guo, Jing Shi, Yen-Ju Lu, Aswin Shanmugam Subramanian, Tianzi Wang, Shu-wen Yang, Yu Tsao, Hung-yi Lee, et al. 2021. An exploration of self-supervised pretrained representations for end-to-end speech recognition. *arXiv preprint arXiv:2110.04590*.

Sanyuan Chen, Chengyi Wang, Zhengyang Chen, Yu Wu, Shujie Liu, Zhuo Chen, Jinyu Li, Naoyuki Kanda, Takuya Yoshioka, Xiong Xiao, et al. 2021. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *arXiv preprint arXiv:2110.13900*.

Chung-Cheng Chiu, Tara N Sainath, Yonghui Wu, Rohit Prabhavalkar, Patrick Nguyen, Zhifeng Chen, Anjuli Kannan, Ron J Weiss, Kanishka Rao, Ekaterina Gonina, et al. 2018. State-of-the-art speech recognition with sequence-to-sequence models. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4774–4778.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *Proc. Interspeech*.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Lenore A Grenoble, Peter K Austin, and Julia Sallabank. 2011. Handbook of endangered languages.

Pengcheng Guo, Florian Boyer, Xuankai Chang, Tomoki Hayashi, Yosuke Higuchi, Hirofumi Inaguma, Naoyuki Kamo, Chenda Li, Daniel Garcia-Romero, Jiatong Shi, et al. 2021. Recent developments on espnet toolkit boosted by conformer. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5874–5878. IEEE.

Hynek Hermansky, Daniel PW Ellis, and Sangita Sharma. 2000. Tandem connectionist feature extraction for conventional hmm systems. In *2000 IEEE international conference on acoustics, speech, and signal processing. Proceedings (Cat. No. 00CH37100)*, volume 3, pages 1635–1638. IEEE.

Chiori Hori, Takaaki Hori, Teng-Yok Lee, Ziming Zhang, Bret Harsham, John R Hershey, Tim K Marks, and Kazuhiko Sumi. 2017. Attention-based multimodal fusion for video description. In *Proceedings of the IEEE international conference on computer vision*, pages 4193–4202.

Wenxin Hou, Yue Dong, Bairong Zhuang, Longfei Yang, Jiatong Shi, and Takahiro Shinozaki. 2020. Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning. In *Proc. Interspeech 2020*, pages 1037–1041.

Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. Hubert: How much can a bad teacher benefit asr pre-training? In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6533–6537. IEEE.

J. Kahn, M. Rivière, W. Zheng, E. Kharitonov, Q. Xu, P. E. Mazaré, J. Karadayi, V. Liptchinsky, R. Collobert, C. Fuegen, T. Likhomanenko, G. Synnaeve, A. Joulin, A. Mohamed, and E. Dupoux. 2020. Libri-light: A benchmark for asr with limited or no supervision. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7669–7673. `https://github.com/facebookresearch/libri-light`.

Anjuli Kannan, Arindrima Datta, Tara N. Sainath, Eugene Weinstein, Bhuvana Ramabhadran, Yonghui Wu, Ankur Bapna, Zhifeng Chen, and Seungji Lee. 2019. Large-Scale Multilingual Speech Recognition with a Streaming End-to-End Model. In *Proc. Interspeech 2019*, pages 2130–2134.

Shigeki Karita, Nanxin Chen, Tomoki Hayashi, Takaaki Hori, Hirofumi Inaguma, Ziyan Jiang, Masao Someki, Nelson Enrique Yalta Soplin, Ryuichi Yamamoto, Xiaofei Wang, et al. 2019. A comparative study on transformer vs rnn in speech applications. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 449–456. IEEE.

Suyoun Kim, Takaaki Hori, and Shinji Watanabe. 2017. Joint ctc-attention based end-to-end speech recognition using multi-task learning. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4835–4839. IEEE.

Partha Lal and Simon King. 2013. Cross-lingual automatic speech recognition using tandem features. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(12):2506–2515.

Bo Li, Ruoming Pang, Tara N Sainath, Anmol Gulati, Yu Zhang, James Qin, Parisa Haghani, W Ronny Huang, Min Ma, and Junwen Bai. 2021. Scaling end-to-end models for large-scale multilingual asr. *arXiv preprint arXiv:2104.14830*.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W Black, and Florian Metze. 2020. Universal phone recognition with a multilingual allophone system.

Jindřich Libovický and Jindřich Helcl. 2017. Attention strategies for multi-source sequence-to-sequence learning. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 196–202, Vancouver, Canada. Association for Computational Linguistics.

Andy T Liu, Shang-Wen Li, and Hung-yi Lee. 2021. Tera: Self-supervised learning of transformer encoder representation for speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2351–2366.

Christoph Lüscher, Eugen Beck, Kazuki Irie, Markus Kitza, Wilfried Michel, Albert Zeyer, Ralf Schlüter, and Hermann Ney. 2019. Rwth asr systems for librispeech: Hybrid vs attention. *Interspeech 2019*.

Krishna D. N, Pinyi Wang, and Bruno Bozza. 2021. Using Large Self-Supervised Models for Low-Resource Speech Recognition. In *Proc. Interspeech 2021*, pages 2436–2440.

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. SpecAugment: A Simple Data Augmentation Method for Automatic Speech Recognition. In *Proc. Interspeech 2019*, pages 2613–2617.

Ngoc-Quan Pham, Thai-Son Nguyen, Jan Niehues, Markus Müller, and Alex Waibel. 2019. Very deep self-attention networks for end-to-end speech recognition. *Proceedings of Interspeech 2019*, pages 66–70.

Vineel Pratap, Anuroop Sriram, Paden Tomasello, Awni Hannun, Vitaliy Liptchinsky, Gabriel Synnaeve, and Ronan Collobert. 2020. Massively Multilingual ASR: 50 Languages, 1 Model, 1 Billion Parameters. In *Proc. Interspeech 2020*, pages 4751–4755.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021a. Leveraging end-to-end ASR for endangered language documentation: An empirical study on yolóxochitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, Online. Association for Computational Linguistics.

Jiatong Shi, Jonathan D Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021b. Highland puebla nahuatl speech translation corpus for endangered language documentation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 53–63.

Shubham Toshniwal, Tara N Sainath, Ron J Weiss, Bo Li, Pedro Moreno, Eugene Weinstein, and Kanishka Rao. 2018. Multilingual speech recognition with a single end-to-end model. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4904–4908. IEEE.

Shinji Watanabe, Takaaki Hori, and John R Hershey. 2017. Language independent end-to-end architecture for joint language identification and speech recognition. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 265–271. IEEE.

Anne Wu, Changhan Wang, Juan Pino, and Jiatao Gu. 2020. Self-Supervised Representations Improve End-to-End Speech Translation. In *Proc. Interspeech 2020*, pages 1491–1495.

Qiantong Xu, Alexei Baevski, and Michael Auli. 2021. Simple and effective zero-shot cross-lingual phoneme recognition.

Brian Yan, Siddharth Dalmia, David R. Mortensen, Florian Metze, and Shinji Watanabe. 2021. Differentiable allophone graphs for language-universal speech recognition. *Proc. Interspeech 2021*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, Tzu-Hsien Huang, Wei-Cheng Tseng, Ko tik Lee, Da-Rong Liu, Zili Huang, Shuyan Dong, Shang-Wen Li, Shinji Watanabe, Abdelrahman Mohamed, and Hung yi Lee. 2021. SUPERB: Speech Processing Universal PERformance Benchmark. In *Proc. Interspeech 2021*, pages 1194–1198.

Cheng Yi, Jianzhong Wang, Ning Cheng, Shiyu Zhou, and Bo Xu. 2020. Applying wav2vec2. 0 to speech recognition in various low-resource languages. *arXiv preprint arXiv:2012.12121*.

Alexander Zahrer, Andrej Zgank, and Barbara Schuppler. 2020. Towards building an automatic transcription system for language documentation: Experiences from muyu. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 2893–2900.

Qifeng Zhu, Barry Chen, Nelson Morgan, and Andreas Stolcke. 2004. On using mlp features in lvcsr. In *Eighth International Conference on Spoken Language Processing*.