INTERSPEECH 2020
October 25–29, 2020, Shanghai, China

Large-Scale End-to-End Multilingual Speech Recognition and Language Identification with Multi-Task Learning

Wenxin Hou[1], *Yue Dong*[1], *Bairong Zhuang*[1], *Longfei Yang*[1], *Jiatong Shi*[2] *and Takahiro Shinozaki*[1]

[1]Tokyo Institute of Technology
[2]Johns Hopkins University

[1]http://www.ts.ip.titech.ac.jp, [2]jiatong_shi@jhu.edu

Abstract

In this paper, we report a large-scale end-to-end language-independent multilingual model for joint automatic speech recognition (ASR) and language identification (LID). This model adopts hybrid CTC/attention architecture and achieves word error rate (WER) of 52.8 and LID accuracy of 93.5 on 42 languages with around 5000 hours of training data. We also compare the effects of using subword-level or character-level vocabulary for large-scale multilingual tasks. Furthermore, we transfer the pre-trained model to 14 low-resource languages. Results show that the pre-trained model achieves significantly better results than non-pretrained baselines on both language-specific and multilingual low-resource ASR tasks in terms of WER, which is reduced by 28.1% and 11.4% respectively.

Index Terms: automatic speech recognition, multilingual, low-resource, transfer learning, language identification

1. Introduction

End-to-end Automatic Speech Recognition (ASR) methods have demonstrated favorable results compared to conventional Hidden Markov Model (HMM) methods [1, 2]. In the previous literature, various architectures for end-to-end ASR using either Connectionist Temporal Classification (CTC) [3, 4] or attention-based encoder-decoder [5, 6]. More recently, Watanabe et al. [7] proposed a hybrid CTC/attention architecture, which benefits from both architectures in training and decoding. Moreover, Nakatani et al. [8] employed the multi-head self-attention mechanism in a similar Transformer architecture [9]. The system realized significant reductions in the word error rate (WER) on several public corpora.

Regardless of the architectures, end-to-end systems generally fold the acoustic, lexicon, and language models into a single network. They save the effort on language-specific processing, making it easier to apply them to new languages. Watanabe et al. [10] first proposed an end-to-end language-independent model for joint multilingual ASR and language identification (LID). They trained a hybrid CTC/attention model on 1327 hours of speech data in ten languages and demonstrated comparable/superior performance to language-dependent end-to-end ASR systems. In addition, models trained in a multilingual manner may share information across languages, which helps improve the performance of low-resource language tasks. Zhou et al. [11] employed Transformer architecture to perform multilingual ASR on low-resource languages and achieved 10% relative improvement on the baseline. Kannan et al. [12] presented a large-scale streaming end-to-end model trained on nine Indian languages using CTC. Cho et al. [13] showed that transfer learning could be adopted for end-to-end multilingual models. In [14, 15, 16, 17], similar empirical results also indicated that multilingual architecture could reduce 10-20% relative WER on low-resource languages.

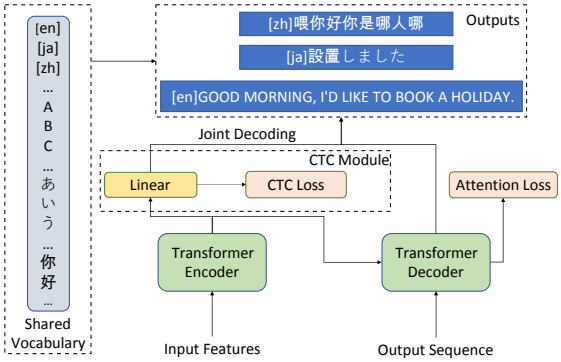

Figure 1: *System architecture*

Inspired by their works, we present this large-scale language-independent multilingual model based on Transformer [9] with hybrid CTC/attention architecture for joint ASR and LID. We increase the number of languages to 42 and the total duration of training data to around 5000 hours with over 6 million utterances. The experiments demonstrate promising results on 42 languages and we obtained significant reductions in WER when transferred to low-resource languages [1].

2. Model Description

This section briefly introduces our Transformer-based language-independent multilingual system with hybrid CTC/attention architecture, which is shown in Figure 1.

2.1. Hybrid CTC/attention Architecture

Similar to [18], we adopt a hybrid CTC/attention architecture, where the model is composed of three components: a shared encoder, an attention decoder, and a CTC module.

Multi-task Learning The hybrid architecture is established in the scheme of multi-task learning. The training process is to jointly optimize the weight-sum of the decoding loss of attention model $\mathcal{L}_{\text{att}}$ and the CTC loss $\mathcal{L}_{\text{ctc}}$ [3]. The multi-task loss function is given by:

$$\mathcal{L} = \alpha\mathcal{L}_{\text{ctc}} + (1-\alpha)\mathcal{L}_{\text{att}}, \quad (1)$$

where hyperparameter α represents the weight of the CTC loss. Previous research [8, 18] has shown that introducing CTC as an auxiliary task can help the model learn appropriate alignments and converge faster.

[1]Recipes for our experiments (without distributed training) are available as part of ESPnet: https://github.com/espnet/espnet

http://dx.doi.org/10.21437/Interspeech.2020-2164

Joint Decoding During decoding, given speech input X, the final prediction is made based on a weighted sum of log probabilities from both the CTC and attention components:

$$\hat{Y} = \underset{Y \in \mathcal{Y}}{\operatorname{argmax}} \{\lambda \log P_{\text{ctc}}(Y|X) + (1-\lambda) \log P_{\text{att}}(Y|X)\}, \quad (2)$$

where λ is a hyperparameter, $P_{\text{ctc}}(Y|X)$ and $P_{\text{att}}(Y|X)$ are the output probabilities of prediction Y from CTC and attention decoder respectively.

2.2. Language-Independent Architecture

For multilingual ASR tasks, we adopt a language-independent architecture [10] so that all the target languages can share the same network architecture and parameters.

Shared Vocabulary The output vocabulary includes characters or subwords of all the target languages. This setting makes it possible to train a single network for all languages in a language-independent manner.

Joint Language Identification To reduce the possibility that predictions switch between languages, we insert the corresponding language ID (e.g., $[en], [fr]$) at the beginning of every output target during training. Consequently, the model can learn to first identify the language and then predict the output text during decoding. This can be regarded as an auxiliary language identification (LID) task.

3. Recognition Experiments

We first train and evaluate quadrant-bi (42) lingual ASR systems. We compare a character and subword unit based modelings. The character vocabulary includes 7,381 characters, and the subword vocabulary includes 15,943 subwords. We tokenized the subwords by using the SentencePiece library [19]. Apart from characters/subwords, 60 non-language symbols such as language IDs and other symbols (e.g., <*UNK*>) are also included in the vocabularies.

We then transfer the character-level 42-lingual model to perform monolingual and quattuordec (14)-lingual ASR on 14 low-resource languages. As the baseline, we train randomly initialized monolingual and 14-lingual models using the low-resource languages.

4. Experimental Setup

4.1. Data

Table 1 shows the data we used for 42-language training and testing. It is from 11 databases including: AISHELL [20], Aurora4, Babel, CHiME4, Common Voice [21], Corpus of Spontaneous Japanese (CSJ) [22], Fisher Switchboard, Fisher Callhome Spanish, HKUST [23], WSJ and Voxforge. For the Common Voice and Voxforge data, we randomly used 80% as a training set, 10% as a development set, and 10% as a test set.

For the low-resource tasks, we select 14 languages from the Common Voice database [21] including Arabic (7 hrs), Breton (5 hrs), Hakha Chin (2 hrs), Chuvash (0.96 hrs), Dhivehi (6 hrs), Esperanto (35 hrs), Estonian (10 hrs), Indonesian (3 hrs), Interlingua (1 hr), Kinyarwanda (0.25 hrs), Kyrgyz (11 hrs), Latvian (4 hrs), Sakha (3 hrs) and Slovenian (3 hrs).

Table 1: *Corpora used for 42-language experiment.*

Language	**Corpora**	**#Utterances**		
		train	dev	test
Amharic	Babel	37k	4k	10k
Assamese	Babel	57k	6k	10k
Bengali	Babel	55k	6k	10k
Catalan	CommonVoice	68k	9k	8k
Cantonese	Babel	72k	8k	10k
Cebuano	Babel	39k	4k	11k
Welsh	CommonVoice	27k	3k	4k
German	Voxforge, CommonVoice	256k	33k	32k
Dholuo	Babel	40k	4k	11k
English	Aurora4, CHiME4, Fisher Switchboard, WSJ, CommonVoice	2,760k	83k	68k
Spanish	Fisher Callhome, Voxforge, CommonVoice	254k	11k	18k
Basque	CommonVoice	26k	3k	3k
Persian	CommonVoice	43k	5k	5k
French	Voxforge, CommonVoice	133k	17k	16k
Georgian	Babel	34k	4k	9k
Guarani	Babel	37k	4k	10k
Haitian	Babel	52k	6k	10k
Igbo	Babel	35k	4k	10k
Italian	Voxforge, CommonVoice	30k	4k	4k
Japanese	CSJ	402k	4k	4k
Javanese	Babel	42k	5k	11k
Kabyle	CommonVoice	145k	18k	18k
Kazakh	Babel	43k	5k	12k
Kurmanji	Babel	42k	5k	11k
Lao	Babel	60k	7k	11k
Lithuanian	Babel	36k	4k	10k
Mongolian	Babel	40k	4k	11k
Dutch	Voxforge	7k	1k	1k
Pashto	Babel	63k	7k	9k
Portuguese	Voxforge	3k	0.4k	0.3k
Russian	Voxforge, CommonVoice	20k	3k	3k
Swahili	Babel	40k	4k	11k
Tagalog	Babel	84k	9k	11k
Tamil	Babel	58k	6k	11k
Telugu	Babel	39k	4k	11k
Tok Pisin	Babel	37k	4k	10k
Tatar	CommonVoice	18k	2k	2k
Turkish	Babel	74k	8k	10k
Vietnamese	Babel	71k	8k	9k
Mandarin (zh-CN)	AISHELL, HKUST	621k	18k	13k
Mandarin (zh-TW)	CommonVoice	33k	4k	4k
Zulu	Babel	55k	6k	11ks
Sum		6,088k	356k	464k

4.2. Implementation Details

For all the experiments, 83-dimensional input features are extracted from the raw speech composed of 80-dimensional filter banks and 3-dimensional pitch features computed every 10 ms over a 25 ms window. The detailed Transformer configuration follows the same setting as the *big model* described in [24]. The models are trained using Adam optimizer with a varying learn-

Table 2: *Training and decoding configurations for the 42 and 14 language datasets.*

Hyperparameters	**42-lang.**	**14-lang.**
Training epochs	100	100
Dropout	0.1	0.1
Learning rate factor k	4.5	1.0
Gradient clipping	5	5
Gradient accumulation	1	2
Batch size	1,280	32
Warmup step	25,000	25,000
CTC loss weight α	0.3	0.3
CTC decoding weight λ	0.5	0.5
Beam size	10	10

ing rate lr strategy:

$$lr = k \cdot d_{\text{model}}^{-0.5} \cdot \min(\text{step}^{-0.5}, \text{step} \cdot \text{warmup\_step}^{-1.5}), \quad (3)$$

where hyperparameter k is the learning rate factor, lr linearly warms up before step reaching warmup_step and decreases proportionally to the inverse square root of step afterward.

We employed the ESPnet toolkit [25] to conduct all experiments. The model training and evaluation were performed using the TSUBAME 3.0 supercomputer [2]. To accelerate the 42-language training, we applied PyTorch distributed communication package [3] to train the model on 10 computing nodes with 40 NVIDIA TESLA P100 GPUs with a total batch size of 1,280. During the training process, the character-level model took around 163 hours and the subword-level model took 222 hours. For the fine-tuning experiments, only one GPU was used. Detailed training and decoding hyperparameters are found in Table 2.

4.3. Evaluation Metrics

We used word error rate (WER) and character error rate (CER) as evaluation metrics. To obtain an average over languages, we weighted the WERs by the amount of test data. In addition, we evaluated language identification (LID) accuracy for the language-independent multilingual models. In the event that there were multiple corpora used for one language, we calculated the weighted average over those corpora.

5. Results

We first compare the 42-language task's WERs/CERs and LID accuracies obtained by models based on character-level and subword-level vocabularies. The results are shown in Table 3. We can see that the large-size subword-level vocabulary improves the model's performance in WER and LID accuracy for 38 and 29 languages, respectively. The average WER was reduced from 52.8 % to 49.6 %, and the averaged CER was reduced from 27.8 % to 27.2 %. Meanwhile, the average LID accuracy is increased from 93.5 to 94.0. These results demonstrate the advantage of introducing subwords.

Next we analyzed the 14 low-resource languages' WER obtained by the pre-trained (knowledge transferred) and non-pretrained (randomly initialized) models fine-tuned for the language-specific (one language at a time) and language-independent multilingual (14 languages together) tasks. The results are shown in Table 4. By comparing the second and fourth columns without the pre-training, we see that multilingual training generally improves the model performance on low-resource languages. The transfer learning was effective for all the 14 languages, and the lowest WERs were obtained by one of the pre-trained models. Seven out of 14 languages got the lowest WER in the language-specific knowledge transfer condition, while other seven languages got the best result in the language-independent condition. This was maybe related to the similarity/dissimilarity between languages.

By applying pre-training, the weighted averages of WER

[2]https://www.gsic.titech.ac.jp/en/tsubame

[3]https://pytorch.org/docs/stable/distributed.html

Table 3: *Word and character error rates (WER/CER) and language identification (LID) accuracy of character-based model (Char.) and subword-baesd model (SubW.). For Cantonese, Japanese, and Mongolian, we used sentence error rate instead of word error rate.*

	WER/CER		LID acc.	
Language	Char.	SubW.	Char,	SubW.
Amharic	61.2/42.6	**57.5/40.1**	92.1	**92.9**
Assamese	67.3/44.9	**61.4/42.9**	81.5	**83.6**
Bengali	65.0/39.5	**60.8/38.5**	**82.9**	81.1
Catalan	40.0/10.6	**36.8/9.7**	99.1	99.1
Cantonese	99.5/66.6	**99.4/65.7**	99.8	99.8
Cebuano	63.9/37.4	**56.8/35.7**	81.5	**83.2**
Welsh	38.4/12.3	**36.4/11.9**	96.5	**97.7**
German	32.5/8.6	**30.8/8.1**	99.3	**99.4**
Dholuo	58.4/32.8	**52.5/31.3**	**87.6**	86.2
English	21.1/8.6	**19.3/8.2**	**99.3**	99.2
Spanish	**50.3/19.0**	50.8/21.0	98.6	**98.9**
Basque	37.0/6.8	**32.9/6.1**	99.4	**99.6**
Persian	47.9/15.0	**44.2/13.6**	99.2	**99.5**
French	44.5/14.2	**40.6/13.0**	99.4	**99.5**
Georgian	63.8/36.8	**58.6/34.7**	86.1	**88.9**
Guarani	82.3/42.6	**78.4/41.0**	82.1	**83.5**
Haitian	63.4/36.5	**58.4/34.8**	92.8	**93.8**
Igbo	71.5/41.8	**65.0/40.4**	**84.3**	84.2
Italian	40.3/9.9	**36.3/9.0**	97.9	**98.6**
Japanese	62.3/7.2	**60.5/6.7**	100.0	100.0
Javanese	71.6/46.3	**65.5/45.2**	80.3	**82.5**
Kabyle	54.7/17.1	**52.3/16.4**	99.7	**99,8**
Kazakh	104.2/**90.5**	**103.9**/90.8	**82.5**	81.7
Kurmanji	88.8/52.7	**85.3/52.4**	**90.0**	89.9
Lao	53.4/37.3	**48.1/34.8**	87.6	**89.6**
Lithuanian	72.2/42.7	**69.3/42.1**	82.6	**84.6**
Mongolian	**102.5/88.7**	103.6/89.3	91.5	**92.2**
Dutch	54.0/16.8	**50.8/16.2**	97.8	**98.8**
Pashto	57.0/35.7	**50.8/33.3**	93.3	**94.7**
Portuguese	74.6/**28.4**	**73.4**/28.6	97.4	**98.2**
Russian	**90.2/65.9**	92.1/66.4	98.8	**99.5**
Swahili	61.4/32.2	**53.7/30.4**	79.1	**84.9**
Tagalog	60.0/38.4	**53.7/37.0**	87.9	**89.5**
Tamil	76.0/46.1	**70.9/43.6**	86.4	**87.6**
Telugu	80.3/50.0	**75.8/47.2**	84.6	**85.3**
Tok Pisin	43.2/28.0	**38.3/26.7**	91.4	**92.8**
Tatar	**102.8**/83.6	103.0/**83.3**	98.8	98.8
Trukish	76.2/38.1	**72.1/37.1**	88.9	**89.3**
Vietnamese	95.7/55.8	**94.9/54.9**	90.2	**91.1**
Mandarin(zh-CN)	68.5/14.7	**65.4/13.5**	99.7	99.7
Mandarin(zh-TW)	76.0/22.5	**74.3/21.8**	99.9	99.9
Zulu	70.7/37.6	**66.9/37.3**	**94.5**	93.9
Weighted Avg.	52.8/27.8	**49.6/27.2**	93.5	**94.0**

Table 4: *Comparison of language-specific and language-independent multilingual experiments using non-pretrained baselines and pre-trained models in word error rate (WER) for 14 low-resource languages*

	Language-specific w/o pre-train	Language-specific w/ pre-train	Language-independent w/o pre-train	Language-independent w/ pre-train
Arabic	88.8	**46.4**	56.5	47.8
Breton	90.7	51.9	61.3	**50.1**
Hakha Chin	86.6	42.6	43.6	**34.2**
Chuvash	148.0	147.2	104.4	**101.8**
Dhivehi	95.1	**54.7**	63.0	55.2
Esperanto	24.9	**12.1**	28.1	23.5
Estonian	91.6	**48.3**	68.2	56.4
Indonesian	89.6	50.0	56.0	**43.9**
Interlingua	107.7	76.6	53.7	**39.0**
Kinyarwanda	220.1	222.8	101.6	**100.5**
Kyrgyz	72.6	**33.0**	68.3	62.9
Latvian	85.7	**40.8**	102.1	93.5
Sakha	99.6	**58.4**	102.9	102.5
Slovenian	85.0	79.1	57.0	**50.4**
Weighted Avg.	83.4	60.0	56.9	**50.4**

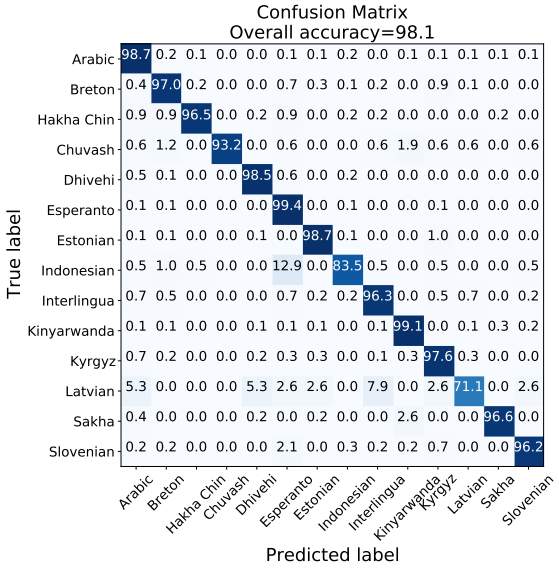

Figure 2: *Confusion matrix of LID accuracies and error rates for the low-resource languages obtained by the non-pretrained language-independent multilingual baseline.*

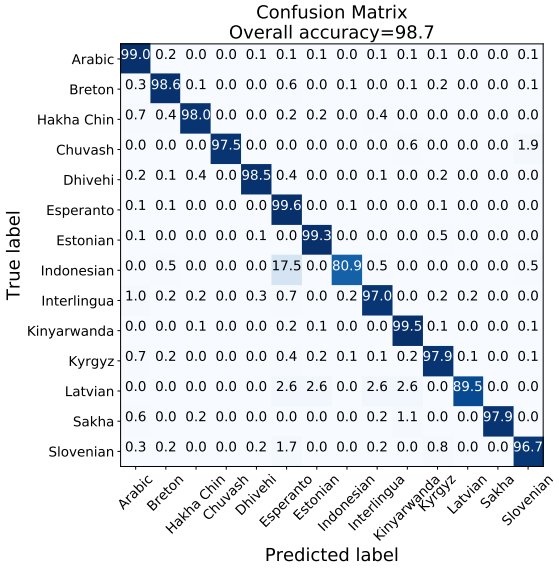

Figure 3: *Confusion matrix of LID accuracies and error rates for the low-resource languages obtained by the pre-trained language-independent multilingual model.*

were reduced by 28.1% and 11.4% from the baselines on language-specific and multilingual tasks, respectively. Finally, Figures 2 and 3 show the confusion matrices of LID accuracies and error rates over 14 low-resource languages that were obtained by the multilingual models with and without pre-training. We can observe that the model with pre-training generally achieves higher accuracy in 13 languages. With these results, we can safely conclude that the pre-trained model efficiently improves the performance on low-resource languages.

6. Conclusions

In this work, we present a large-scale end-to-end language-independent multilingual model for joint ASR and LID based on a transformer-incorporating hybrid CTC/attention architecture that is trained with up to 6 million utterances. Our results are promising with an average WER of 52.8 and an average LID accuracy of 93.5 for 42 languages. We also find that using a large-size subword-level vocabulary can further improve the model's performance in multilingual tasks. Moreover, we show the significant improvement that large-scale pre-training brings about in the model's performance on low-resource languages.

Our future work will include handling the imbalanced training data in order to utilize data more efficiently. We also will investigate the effects of linguistic connections (e.g., language families) between languages in order to further improve multilingual ASR and LID.

7. Acknowledgements

We would like to thank Prof. Shinji Watanabe from Johns Hopkins University for his assistance with this project.

8. References

[1] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International conference on machine learning*, 2014, pp. 1764–1772.

[2] W. Han, Z. Zhang, Y. Zhang, J. Yu, C.-C. Chiu, J. Qin, A. Gulati, R. Pang, and Y. Wu, "Contextnet: Improving convolutional neural networks for automatic speech recognition with global context," *arXiv preprint arXiv:2005.03191*, 2020.

[3] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber, "Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks," in *Proceedings of the 23rd international conference on Machine learning*, 2006, pp. 369–376.

[4] Y. Miao, M. Gowayyed, and F. Metze, "Eesen: End-to-end speech recognition using deep rnn models and wfst-based decoding," in *2015 IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*. IEEE, 2015, pp. 167–174.

[5] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in neural information processing systems*, 2015, pp. 577–585.

[6] W. Chan, N. Jaitly, Q. Le, and O. Vinyals, "Listen, attend and spell: A neural network for large vocabulary conversational speech recognition," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2016, pp. 4960–4964.

[7] S. Watanabe, T. Hori, S. Kim, J. R. Hershey, and T. Hayashi, "Hybrid ctc/attention architecture for end-to-end speech recognition," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1240–1253, 2017.

[8] T. Nakatani, "Improving transformer-based end-to-end speech recognition with connectionist temporal classification and language model integration," 2019.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[10] S. Watanabe, T. Hori, and J. R. Hershey, "Language independent end-to-end architecture for joint language identification and speech recognition," in *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2017, pp. 265–271.

[11] S. Zhou, S. Xu, and B. Xu, "Multilingual end-to-end speech recognition with a single transformer on low-resource languages," *arXiv preprint arXiv:1806.05059*, 2018.

[12] 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," *arXiv preprint arXiv:1909.05330*, 2019.

[13] 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 *2018 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2018, pp. 521–527.

[14] S. Toshniwal, T. N. Sainath, R. J. Weiss, B. Li, P. Moreno, E. Weinstein, and K. Rao, "Multilingual speech recognition with a single end-to-end model," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4904–4908.

[15] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 4909–4913.

[16] T. Sercu, G. Saon, J. Cui, X. Cui, B. Ramabhadran, B. Kingsbury, and A. Sethy, "Network architectures for multilingual speech representation learning," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 5295–5299.

[17] A. Datta, B. Ramabhadran, J. Emond, A. Kannan, and B. Roark, "Language-agnostic multilingual modeling," *arXiv preprint arXiv:2004.09571*, 2020.

[18] S. Kim, T. Hori, and S. Watanabe, "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)*. IEEE, 2017, pp. 4835–4839.

[19] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," *arXiv preprint arXiv:1804.10959*, 2018.

[20] H. Bu, J. Du, X. Na, B. Wu, and H. Zheng, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *2017 20th Conference of the Oriental Chapter of the International Coordinating Committee on Speech Databases and Speech I/O Systems and Assessment (O-COCOSDA)*. IEEE, 2017, pp. 1–5.

[21] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[22] K. Maekawa, "Corpus of spontaneous japanese: Its design and evaluation," in *ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition*, 2003.

[23] Y. Liu, P. Fung, Y. Yang, C. Cieri, S. Huang, and D. Graff, "Hkust/mts: A very large scale mandarin telephone speech corpus," in *International Symposium on Chinese Spoken Language Processing*. Springer, 2006, pp. 724–735.

[24] L. Dong, S. Xu, and B. Xu, "Speech-transformer: a no-recurrence sequence-to-sequence model for speech recognition," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 5884–5888.

[25] S. Watanabe, T. Hori, S. Karita, T. Hayashi, J. Nishitoba, Y. Unno, N. Enrique Yalta Soplin, J. Heymann, M. Wiesner, N. Chen, A. Renduchintala, and T. Ochiai, "Espnet: End-to-end speech processing toolkit," in *Interspeech*, 2018, pp. 2207–2211. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2018-1456