INTERSPEECH 2021
30 August – 3 September, 2021, Brno, Czechia

Language and Speaker-Independent Feature Transformation for End-to-End Multilingual Speech Recognition

Tomoaki Hayakawa[1], *Chee Siang Leow*[1], *Akio Kobayashi*[2], *Takehito Utsuro*[3], *Hiromitsu Nishizaki*[1]

[1]Graduate School of Medicine, Engineering, and Agricultural Sciences, University of Yamanashi, Japan
[2]Faculty of Industrial Technology, Tsukuba University of Technology, Japan
[3]Faculty of Engineering, Information and Systems, University of Tsukuba, Japan

{kurotomo,cheesiang_leow}@alps-lab.org, a-kobayashi@a.tsukuba-tech.ac.jp, utsuro@iit.tsukuba.ac.jp, hnishi@yamanashi.ac.jp

Abstract

This paper proposes a method to improve the performance of multilingual automatic speech recognition (ASR) systems through language- and speaker-independent feature transformation in a framework of end-to-end (E2E) ASR. Specifically, we propose a multi-task training method that combines a language recognizer and a speaker recognizer with an E2E ASR system based on connectionist temporal classification (CTC) loss functions. We introduce the language and speaker recognition sub-tasks into the E2E ASR network and introduce a gradient reversal layer (GRL) for each sub-task to achieve language and speaker-independent feature transformation. The evaluation results of the proposed method in the multilingual ASR system in six sorts of languages show that the proposed method achieves higher accuracy than the ASR models for each language by introducing multi-tasking and GRL.

Index Terms: end-to-end multilingual ASR, feature transformation, gradient reversal layer, multi-task training

1. Introduction

The performance of speech recognition systems has improved dramatically with the introduction of deep learning techniques. In recent years, end-to-end (E2E) automatic speech recognition (ASR) systems such as ESPNet [1] and Deep Speech [2, 3] have attracted much attention. They are approaching the performance of Deep Neural Network-Hidden Markov Model (DNN-HMM)-based ASR systems such as Kaldi [4].

Most ASR research targets only single language speech recognition. However, there is also much research on low-resource language ASR that utilizes speech data from languages different than the target language rather than only the target language data for lowresource. They are used to compensate for the lack of training data on low-resource languages by using data from richer languages. Thus, there are many examples of research on ASR for a single language using multiple language resources [5, 6].

On the other hand, there is a demand for a speech recognition system that can handle multiple languages simultaneously in a globalized society. In the past, methods that use a language recognition technique have been proposed [7, 8]. Following language recognition, a language-specific ASR system was adapted to an input speech. However, in this approach, a language recognition system is needed. In addition, in the case of a code-switched word included in an utterance, it does not work well. Therefore, a single speech recognition system that can recognize multiple languages simultaneously is desirable; however, there are still a few research examples. As an example, Seki et al. [9] proposed a multilingual ASR that uses language recognition from the first part of an utterance. There is also an increasing number of studies dealing with code-switching in which words from more than one language are mixed in with an utterance [10, 11, 12].

We have already proposed a multilingual ASR method [13]. In this method, we defined a shared phoneme set for six sorts of languages using the International Phonetic Alphabet (IPA) [14], and the acoustic model for multilingual ASR was then trained with the shared phoneme set. As a result, the ASR performances have improved compared to the ASR model trained from a single language resource.

In this study, we propose a method to improve the accuracy of multilingual ASR using the E2E ASR system framework. Specifically, we propose a multi-task training method that combines a language and speaker recognizer with the E2E model based on connectionist temporal classification (CTC) [15] loss functions. The aim of introducing these sub-tasks is to transform acoustic features that are invariant to language and speaker. We introduce a framework of back-propagation to the neural network for ASR to achieve language- and speaker-independent feature transformation by using a gradient reversal layer (GRL) [16, 17]. The multi-task learning research on multi-language ASR systems conducted by Huo et al. [18] and Cho et al. [19] aims to improve the robustness of multi-language ASR by combining S2S with Attention and CTC loss. On the other hand, our experiment aims to extract language-independent universal features by using multiple identification criteria, i.e., language identification and speaker identification, together, and the purpose and aim of our research are completely different from papers [18] and [19]. Although there are individual studies on universalization and language identification using GRL, there have been no studies that have applied GRL to multiple language ASR, and in this respect, this study is original.

The target languages in this study are Czech, English, French, German, Japanese, and Spanish. Although these are not generally low-resource languages, we limit the training data for each language to about 17 hours and show that speech recognition is possible even with a small amount of data. There are some studies on the use of GRL for ASR [20]. Denisov et al. used the GRL-based speaker recognition model to resolve mismatch between recording environments. Yi et al. [21] proposed adversarial multilingual training to train neural networks for a specific language. This study is close to our work. However, our study combines a language and speaker recognition network with the GRL. The contributions of this research are as follows:

- To improve the accuracy of ASR in E2E ASR by com-

http://dx.doi.org/10.21437/Interspeech.2021-390

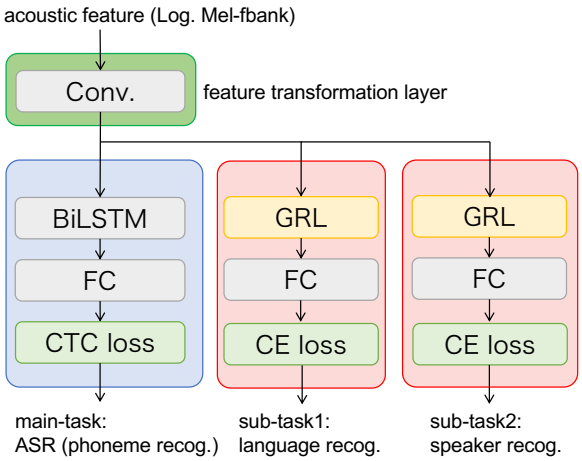

Figure 1: *E2E ASR architecture with multi-task training*

bining a language and a speaker recognizer.

- The combination of a language and speaker recognizer improves the ASR accuracy of multilingual speech.
- It is possible to improve the E2E ASR model by using multilingual data simultaneously.

The experimental results showed that our E2E ASR model, which was trained under the multi-task training framework with the language and speaker recognition models with the GRL, outperformed the ASR model trained on data from a single language. Therefore, the GRL could realize language- and speaker-independent feature transformation, and it was effective in multilingual ASR.

2. Multi-task E2E ASR model

2.1. End-to-End ASR Model

Figure 1 shows an architecture of our E2E ASR model in which the main-task (ASR) network and the sub-task (language and speaker recognition) networks are combined. The whole E2E ASR model is composed of the Conv. layers, which can translate the acoustic feature input to the network into the feature representations. The output of the Conv. layers is fed into the ASR network and the language (speaker) recognition network. The sub-task networks include the GRL, which affects the training of the parameters in the Conv. layers. The ASR (main-task) network consists of four BiLSTM layers and an FC layer without any attention mechanisms. This is because we would just like to investigate the effectiveness of the GRL on multilingual ASR.

2.2. Gradient Reversal Layer

As shown in Figure 2, the GRL is a layer in which the value of the gradient is multiplied by λ when the gradient is back propagated. By setting λ to a negative value, the gradient is reversed and back propagated in the Conv. layers. Therefore, the Conv. layers are trained to output feature representations in which it is difficult to discriminate between language or speaker. At the same time, the discrimination capability of the sub-network is enhanced in all the fully connected layers of the sub-task network. The GRL can realize language and speaker-independent feature transformation in the Conv. layers, which is expected to facilitate the learning of a multilingual ASR network.

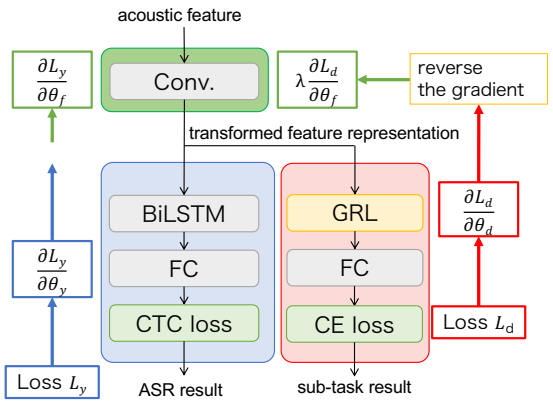

Figure 2: *Back propagation based on the GRL on our E2E ASR network*

Table 1: *The Corpora statistics. The numbers in parentheses mean the number of speakers*

Language	training		test	
	hours	#utt. (#spk)	hours	#utt. (#spk)
Czech	17	6,371 (46)	2.7	1,030 (10)
English	17	7,329 (126)	2.6	1,155 (11)
French	17	6,562 (63)	2.0	821 (8)
German	17	9,258 (71)	1.5	826 (6)
Japanese	17	11,843 (17)	5.1	3,946 (30)
Spanish	17	5,290 (81)	1.7	564 (8)

The value of λ is scheduled by the following equation:

$$\lambda(p) = \frac{2}{1 + \exp(-10p)} - 1$$

, where p is the rate of learning progress. For example, p is set to 0.1 at the initial training epoch if the number of epochs is scheduled to 10.

3. Experiments

3.1. Experimental Setup

Table 1 shows the statistics of the datasets used in this paper. We prepared six sorts of language speech corpora as follows:

- Spanish, Czech, German and French from the GlobalPhone corpus [22],
- English from the TED corpus [23],
- Japanese from the Corpus of Spontaneous Japanese [24].

To ensure a low-resource training condition, we limit the data size for each language to 17 hours because the GlobalPhone corpus for German includes only 17 hours of speech data.

Figure 3 shows the detailed network design of our E2E ASR model with the sub-task network. In this paper, we try to train five sorts of ASR models as follows:

Baseline : only use the main-task network trained by single language data.

CTC pre-train : only use the main-task network trained by the mixed-language data.

ASR+LR : a combination of the main-task and the language recognition networks.

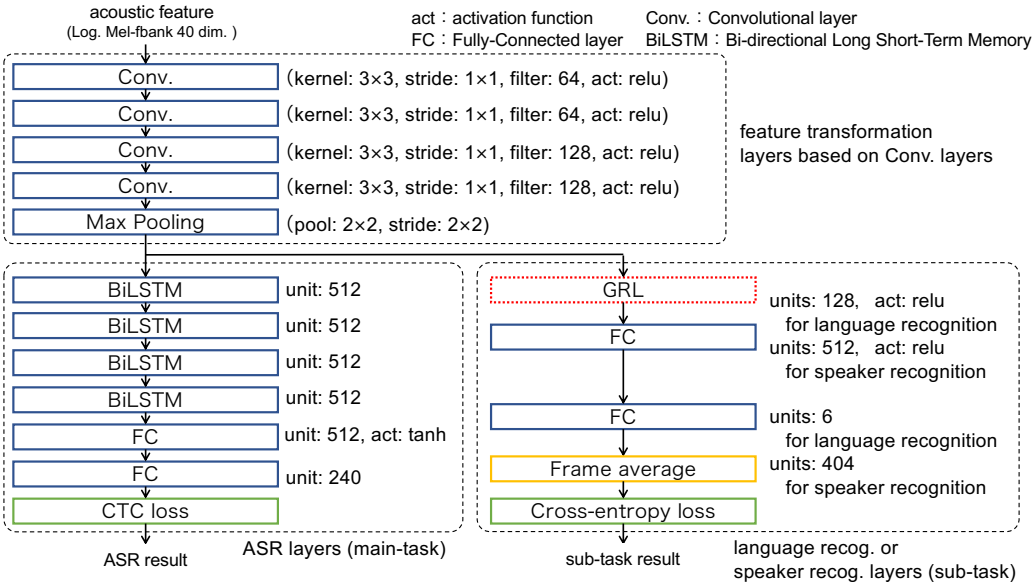

Figure 3: *Detailed design of our end-to-end ASR model*

ASR+SR : a combination of the main-task and the speaker recognition networks.

ASR+LR+SR : a combination of the main-task and the two sub-tasks' networks.

In addition, we also train the combination models in both of with the GRL and without the GRL.

Table 2 shows the hyper-parameters for the multi-task training. The whole loss value of the combined model is calculated as:

$$Loss = Loss_{CTC} + \beta \times Loss_{sub}$$

β are the weight adjustment factors. β is set to 20.0 and 5.0 for the language recognition task and the speaker recognition task, respectively. For the model with the two sub-tasks (ASR+LR+SR), we calculate the whole loss value as follows:

$$Loss = Loss_{CTC} + 10.0 \times Loss_{LR} + 2.5 \times Loss_{SR}$$

These factors are decided by using the validation set.

In this paper, the main-task is phoneme-based ASR. Therefore, we used phoneme error rate (PER) as an evaluation measure. The reason for using PER is that we would like to investigate and analyze the effect of phoneme confusion (confusion errors) across languages. For this investigation, it is desirable to have conditions where the difference in the number of types of output symbols per language is small. For example, in Japanese, which includes Kanji characters, the number of output nodes is large (several thousand or more), and the frequency of occurrence of each character type is highly biased, especially when the corpus is small. It would significantly affect recognition accuracy and may prevent accurate phoneme confusion error analysis. Therefore, in this paper, we used PER as an evaluation measure. The number of phonemes for each language is around 40. Therefore, a total of 238 sorts of phonemes are used in this experiment. We use a phoneme-based trigram language model for decording (beam-search) the output posterior-gram sequences output by the CTC model.

Table 2: *Hyper-parameters for the E2E ASR model training*

Feature	Log Mel-fbank (40 dim.)
Num. of epochs	20 (pre-train), 10 (multi-task)
Mini-batchsize	10 utterances
Optimization	Adam (lr:0.0001, beta1:0.9, beta2:0.999)
Dropout ratio	0.2 for each layer
Loss func.	CTC (pre-train) CTC + cross entropy (multi-task)

3.2. Pre-training

Because our E2E ASR model (including main-task and sub-task network) cannot be trained from scratch using all the speech data from all languages, the only main-task network is pre-trained by following the training schedule:

1 to 3 epochs: only use Spanish utterances

4 to 6 epochs: Spanish and English utterances

7 to 9 epochs: Spanish, English and German utterances

10 to 12 epochs: Spanish, English, German and French utterances

13 to 15 epochs: All languages except Japanese

15 to 20 epochs: All languages

The training order is very important. When the model is trained from scratch, mixed-language data usage cannot train the model well in the first training process (the loss value diverges). Therefore, we start the training from a single language (Spanish) and then we gradually increase the languages. Table 2 also shows the pre-training conditions for the E2E ASR model. After the main-task network training has been accomplished, the whole network, including the sub-task network, is trained from all the language speech data.

3.3. Results

Table 3 shows the phoneme error rates for each training condition. The baseline model, which is trained from a single lan-

Table 3: *Phoneme Error Rates (PERs) [%] for each language*

Language	Baseline		CTC+LR		CTC+SR		CTC+LR+SR	
		CTC pre-train	w/o GRL	w/ GRL	w/o GRL	w/ GRL	w/o GRL	w/ GRL
Czech	15.6	15.3	13.1	13.1	13.3	**12.9**	13.0	**12.9**
English	26.8	27.4	24.1	23.6	24.7	24.0	24.1	**23.0**
Franch	13.5	16.6	12.5	12.8	13.1	12.9	12.6	**12.5**
German	21.0	22.2	19.3	19.3	20.2	19.5	19.6	**19.0**
Japanese	17.8	20.1	17.2	17.1	18.0	17.0	17.3	**16.9**
Spanish	13.2	13.3	11.3	11.7	12.3	11.7	11.9	**11.2**
All languages	18.0	19.4	16.5	16.4	17.1	16.5	16.6	**16.1**

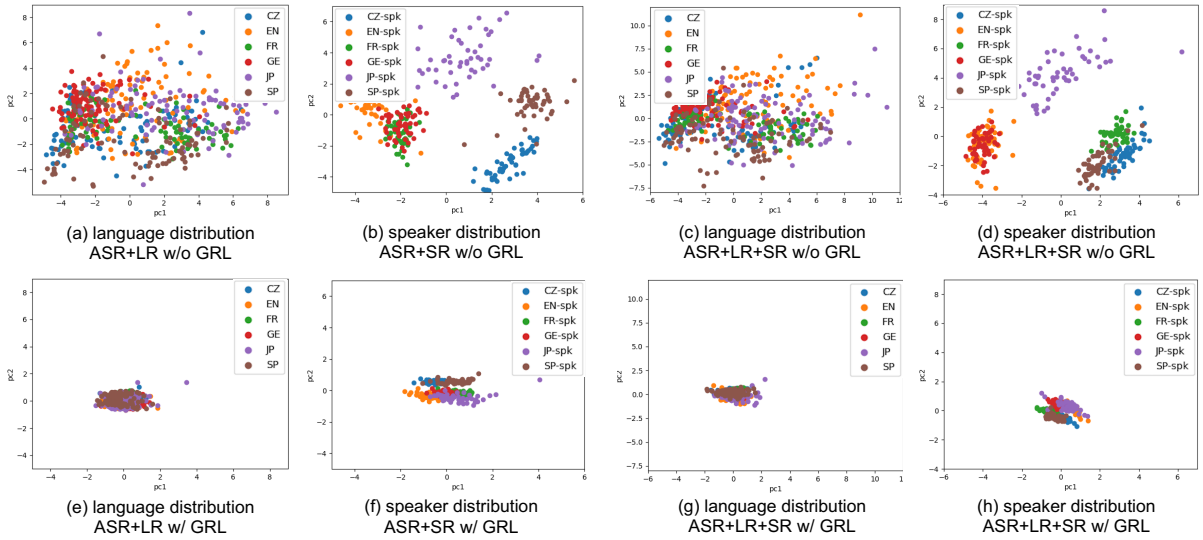

Figure 4: *Distributions of utterance representations (output of the feature transformation layer) for each language*

guage, was better with most language data than the pre-trained CTC model trained from the mixed-language data. Therefore, we could not build a multilingual ASR system simply by mixing multiple language data under the low-resource condition (each set of language data was limited to 17 hours). However, after introducing both sub-tasks without the GRL to the main ASR network, most of the PERs improved compared to the baselines. Moreover, when the GRL was adapted to each sub-task, the PERs further improved compared to without the GRL model. This indicated that the GRL provides speaker- and language-independent feature transformation in the Conv. layers. The best PER result was achieved when both sub-tasks (LR+SR) were introduced. Therefore, removing language and speaker dependencies is effective in multilingual ASR.

Figure 4 shows the distribution of utterance representations for each language and speaker. After an utterance is input to the Conv. layers (feature transformation layer), we take the average of all the intermediate vectors output from the Conv. layers. The averaged vector is then compressed to the two-dimensional vector by using principal component analysis (PCA). Each compressed vector is plotted. As shown in Figure 4, the utterance vectors are widely distributed when the GRL is not used in the sub-task network. For example, in the case of the ASR+LR network, the vectors are nicely distributed for each language. On the other hand, after introducing the GRL to the network, the utterance vectors are denser than without the GRL. Language- and speaker-independent feature transformation by the GRL could extract common features between languages, and the transformed features complemented each other and made the multilanguage ASR model more robust.

4. Conclusions

In this paper, we proposed a GRL-based language and speaker-independent feature transformation to improve the accuracy of multilingual ASR. The E2E ASR network was multi-task trained with a language recognizer with the GRL, a speaker recognizer with the GRL, or a combination of both recognizers. The experimental results showed that the log mel-fbank features input to the E2E network were generalized regardless of language and speaker by the GRL, and that it contributed to improving the accuracy of multilingual ASR.

In the future, we will confirm the effectiveness of the GRL on more complex models with an attention mechanism, and we will also use other sub-tasks to further improve its accuracy.

5. Acknowledgements

This work was supported by JSPS KAKENHI Grant Number 21H00901. Besides, a part of this work was also supported by the Hoso Bunka Foundation.

6. References

[1] Shinji Watanabe, Takaaki Hori, Shigeki Karita, Tomoki Hayashi, Jiro Nishitoba, Yuya Unno, Nelson Enrique Yalta Soplin, Jahn Heymann, Matthew Wiesner, Nanxin Chen, Adithya Renduchintala, and Tsubasa Ochiai, "ESPNet: End-to-end speech processing toolkit," in *Proceedings of INTERSPEECH2018*, 2018, pp. 2207–2211.

[2] Awni Y. Hannun, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, and Andrew Y. Ng, "Deep speech: Scaling up end-to-end speech recognition," *ArXiv preprint, 1412.5567*, 2014.

[3] Dario Amodei et al., "Deep speech 2: end-to-end speech recognition in english and mandarin," in *Proceedings of the 33rd International Conference on International Conference on Machine Learning (ICML'16)*. 2016, pp. 173–182.

[4] Daniel Povey, Arnab Ghoshal, Gilles Boulianne, Lukas Burget, Ondrej Glembek, Nagendra Goel, Mirko Hannemann, Petr Motlicek, Yanmin Qian, Petr Schwarz, Jan Silovsky, Georg Stemmer, and Karel Vesely, "The Kaldi Speech Recognition Toolkit," in *IEEE 2011 Workshop on Automatic Speech Recognition and Understanding (ASRU 2011)*. 2011.

[5] Enno Hermann and Sharon Goldwater, "Multilingual bottleneck features for subword modeling in zero-resource languages," in *Proceedings of INTERSPEECH2018*, 2018, pp. 2668–2672.

[6] Manjunath K E, K. Sreenivasa Rao, Dinesh Babu Jayagopi, and V Ramasubramanian, "Indian languages ASR: A multilingual phone recognition framework with IPA based common phone-set, predicted articulatory features and feature fusion," in *Proceedings of INTERSPEECH2018*, 2018, pp. 1016–1020.

[7] W. Geng, J. Li, Shanshan Zhang, Xinyuan Cai, and Bo Xu, "Multilingual tandem bottleneck feature for language identification," in *Proceedings of INTERSPEECH2015*, 2015, pp. 413–417.

[8] Hwamin Kim and Jeong-Sik Park, "Automatic language identification using speech rhythm features for multi-lingual speech recognition," *Applied Sciences*, vol. 10, no. 7, pp. 2225, 2020.

[9] H. Seki, S. Watanabe, T. Hori, J. L. Roux, and J. R. Hershey, "An end-to-end language-tracking speech recognizer for mixed-language speech," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4919–4923.

[10] E. Yılmaz, H. van den Heuvel, and D. van Leeuwen, "Code-switching detection using multilingual dnns," in *Proceedings of the 2016 IEEE Spoken Language Technology Workshop (SLT)*, 2016, pp. 610–616.

[11] Zhiping Zeng, Yerbolat Khassanov, Van Tung Pham, Haihua Xu, Eng Siong Chng, and Haizhou Li, "On the end-to-end solution to mandarin-english code-switching speech recognition," in *Proceedings of INTERSPEECH2019*, 2019, pp. 2165–2169.

[12] Ne Luo, Dongwei Jiang, Shuaijiang Zhao, Caixia Gong, Wei Zou, and Xiangang Li, "Towards end-to-end code-switching speech recognition," *ArXiv preprint, 1810.13091*, 2018.

[13] S. Hara and H. Nishizaki, "Acoustic modeling with a shared phoneme set for multilingual speech recognition without code-switching," in *Proceedings of the 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2017, pp. 1617–1620.

[14] The IPA 1989 Kiel Convention, "The IPA 1989 kiel convention," *Journal of the International Phonetic Association*, vol. 19, pp. 67–82, 1989.

[15] Alex Graves, Santiago Fernández, Faustino Gomez, , and Jürgen Schmidhuber, "Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks," in *Proceedings of the 23rd International Conference on Machine Learning (ICML'06)*, 2006, pp. 369–376.

[16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky, "Domain-adversarial training of neural networks," *Journal of Machine Learning Research*, vol. 17, no. 1, pp. 2096–2030, 2016.

[17] Yaroslav Ganin and Victor Lempitsky, "Unsupervised domain adaptation by backpropagation," in *Proceedings of Machine Learning Research*, 2015, pp. 1180–1189.

[18] 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 *Proceedings of INTERSPEECH2020*, 2020, pp. 1037–1041.

[19] J. Cho, M. K. Baskar, R. Li, M. Wiesner, S. H. Mallidi, N. Yalta, M. Karafiát, S. Watanabe, and T. Hori, "Multilingual sequence-to- series speech recognition: architecture, transfer learning, and lan- guage modeling," in *Proceedings of the 2018 IEEE Spoken Language Technology Workshop (SLT)*, 2014, pp. 521–427.

[20] Pavel Denisov, Ngoc Thang Vu, and M. Ferras, "Unsupervised domain adaptation by adversarial learning for robust speech recognition," *ArXiv preprint, 1807.11284*, 2018.

[21] J. Yi, J. Tao, Z. Wen, and Y. Bai, "Adversarial multilingual training for low-resource speech recognition," in *Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2018, pp. 4899–4903.

[22] Tanja Schultz, "GlobalPhone: a multilingual speech and text database developed at karlsruhe university," in *Proceedings of ICSLP2002*, 2002, pp. 345–348.

[23] Anthony Rousseau, Paul Deléglise, and Yannick Estève, "Tedlium: an automatic speech recognition dedicated corpus," in *Proceedings of the Conference on Language Resources and Evaluation (LREC2012)*, 2012, pp. 125–129.

[24] K. Maekawa, "Corpus of Spontaneous Japanese: Its design and evaluation," in *Proceedings of the ISCA & IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003)*. 2003, pp. 7–12.