Interspeech 2018
2-6 September 2018, Hyderabad, India

Fast Language Adaptation Using Phonological Information

Sibo Tong[1,2], *Philip N. Garner*[1], *Hervé Bourlard*[1,2]

[1]Idiap Research Institute, Martigny, Switzerland
[2]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland
{sibo.tong, phil.garner, bourlard}@idiap.ch

Abstract

Phoneme-based multilingual connectionist temporal classification (CTC) model is easily extensible to a new language by concatenating parameters of the new phonemes to the output layer. In the present paper, we improve cross-lingual adaptation in the context of phoneme-based CTC models by using phonological information. A universal (IPA) phoneme classifier is first trained on phonological features generated from a phonological attribute detector. When adapting the multilingual CTC to a new, never seen, language, phonological attributes of the unseen phonemes are derived based on phonology and fed into the phoneme classifier. Posteriors given by the classifier are used to initialize the parameters of the unseen phonemes when extending the multilingual CTC output layer to the target language. Adaptation experiments show that the proposed initialization approaches further improve the cross-lingual adaptation on CTC models and yield significant improvements over Deep Neural Network / Hidden Markov Model (DNN/HMM)-based adaptation using limited data.

Index Terms: crosslingual adaptation, connectionist temporal classification (CTC), phonological features, DNN-based speech recognition

1. Introduction

Fast bootstrapping of new languages remains a challenge in the ASR community. A common approach for creating models for low-resourced languages is to transfer the knowledge learned from other well-resourced languages to the target language. For instance, the bottleneck approach aims to extract language-independent phonetic knowledge from a bottleneck layer of a multilingual model and uses bottleneck features as additional inputs to train the acoustic model of a target language [1, 2, 3]. Knowledge can also be transferred by replacing the output layer of a well-trained model and re-training the model to predict the targets of a low-resourced language [4, 5]. All of these approaches are based on a conventional DNN/HMM framework [6, 7]. In order to perform well, DNNs model context-dependent states. However, this creates more challenges for cross-lingual ASR because of the large increase in context-dependent labels arising from the phone set mismatch. When adaptation data is extremely scarce, the performance degradation is still significant.

Recently, the Connectionist Temporal Classification (CTC) framework has been successful in ASR [8]. CTC based systems learn to model context implicitly by the use of a recurrent neural network (RNN). Even monophone-based CTC systems can achieve equal or better performance than DNN/HMM hybrid systems when a large amount of data is available [9, 10]. A phoneme-based CTC model is fundamentally independent of the problem of context-dependent state mismatch, and does not require prior alignments between the input and output, potentially making the cross-lingual adaptation simpler.

Adaptation from an International Phonetic Alphabet (IPA) phoneme-based multilingual CTC model to a low-resourced language has been investigated in [11]. By retaining the parameters already learned in the multilingual output layer and extending it to cover the unseen phonemes, the authors show significant improvements on limited adaptation data. However, the parameters connecting to the unseen phonemes are randomly initialized. The limitations of data-driven approaches appear when training data is limited. Prior human knowledge can help alleviate such bottlenecks. We hypothesize that a better initialization that integrates more human expertise can bootstrap the target network using even less data.

To this end, we start from the IPA universal phoneme-based multilingual CTC model following [11], which is described in Section 2 and is used as our multilingual seed model for cross-lingual adaptation. It has been demonstrated that phonological attribute is a common knowledge source that is fundamental, sharable across all languages and can be properly modelled [12, 13], thus making it attractive for multilingual ASR. However in this paper, we propose to incorporate phonological information to improve the parameter initialization of the unseen phonemes when extending the CTC output layer. Inspired by the Automatic Speech Attribute Transcription (ASAT) framework [14], we first train a phonological attribute detector that detects a collection of phonological attribute cues, and then integrate such cues to make predictions of the same IPA-based multilingual phoneme targets. When a new language arrives, the corresponding phonological attributes of unseen phonemes can be derived from phonological rules. The posteriors produced by the multilingual phoneme classifier indicate how close an unseen phoneme is to those already seen phonemes and can be used for parameter initialization in the CTC model. This proposed approach is presented in Section 3. Experimental results and analysis are provided in Section 4. Finally, Section 5 concludes the paper.

2. Universal Phoneme-based Multilingual CTC Model

Very recently, building end-to-end multilingual speech recognition systems using a universal grapheme set has been investigated [15, 16]. However, modelling graphemes includes implicit modelling of spelling, which requires a large amount of data. Moreover, graphemes can differ a lot from language to language. Languages that have nothing in common in terms of graphemes also share some common phonemes. Moreover, a universal phoneme-based model is easily extensible to unseen phonemes when adapted to a new language.

With this motivation, and following [11, 17], we propose a multilingual architecture that uses a universal output label set consisting of the union of all phonemes from the multiple lan-

10.21437/Interspeech.2018-1990

guages. In this study, the monolingual phones are considered to be the same and merged if they share the same symbol in the IPA table. The network is trained to model the universal phoneme targets using the CTC loss function [8] on data from multiple languages.

In many of our preliminary experiments with CTC, consistent overfitting was observed on limited data. Dropout has been well established for feedforward networks [18]. More recently, various approaches of dropout on feedforward and recurrent connections were explored in the context of CTC [19]. In this work, the same dropout approach, as described in [11, 19], is applied in multilingual training and cross-lingual adaptation to minimize overfitting on limited data.

3. Cross-lingual Adaptation

3.1. Previous Work

The basic procedure of cross-lingual model adaptation on CTC models is simple. As first proposed for DNN models [4], the output layer is removed and a new randomly initialized softmax layer, corresponding to the target language phone set, is added on top of the hidden layers. Usually, the hidden layers are fixed and only the softmax layer will be re-estimated using training data from the target language. If enough data is available, further tuning of the entire network can be considered. Both have been shown effective in [11, 20].

One major advantage of the universal phoneme-based multilingual CTC model over the multilingual DNN is that monophone modeling gets around the problem of mismatch of context-dependent states. When a new target language arrives, it therefore becomes straightforward to extend the existing multilingual model to extra phonemes, rather than discarding all the information already learned in multilingual training. The output layer can be extended by adding connections to the unseen monophones of the target language. Those parameters connecting to the unseen phones are randomly initialized and trained from scratch. The others can be quickly adapted from the multilingual model with little adaptation data. This approach has been proved to be effective on limited adaptation data in our previous work [11], as shown in Figure 1.

When data becomes more scarce, human expertise can be incorporated to boost the model. Phonological studies suggest that each sound unit of a language (phoneme) can be decomposed into a set of phonological features based on the articulators used to produce the sound; the phonological attributes are sharable across all languages. In this paper, we focus on better initialization of the new parameters by using a phonological attribute-based phoneme classifier.

3.2. Phonological Feature-based Phoneme classifier

As shown in Figure 2, the proposed phoneme classifier consists of two main blocks: 1) a data-driven phonological attribute detector, and 2) a frame-based phoneme classifier using phonological features generated from the previous detector.

The phonological attribute detector is a multitask-learning DNN for joint estimation of phonological features. Estimating different phonological features from the same acoustic signal can be considered as a set of interrelated tasks; it has been shown effective for articulatory feature estimation in [21]. To estimate the DNN parameters, multilingual training data is used. The labels for every phonological class are generated from the phoneme alignment according to the phonological

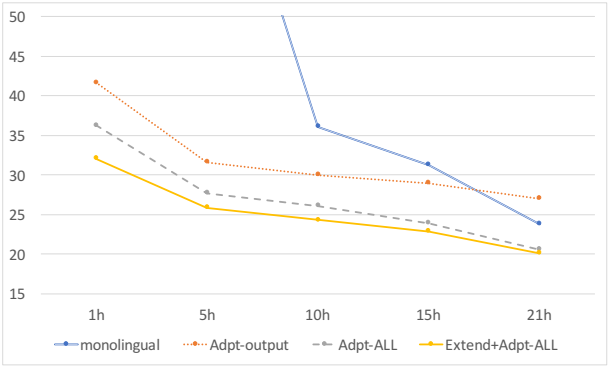

Figure 1: *WERs (%) of different adaptation approaches. Adapt-output denotes only updating the output layer during adaptation. Adpt-ALL is updating the whole network. Extend-Adpt-ALL represents updating the whole network after extending the output layer to the new language. (Figure from [11]).*

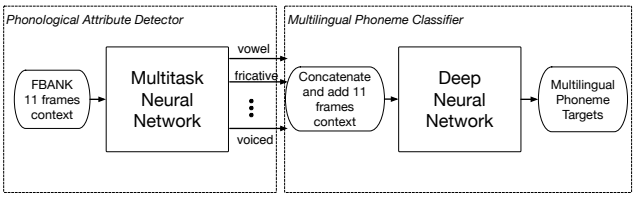

Figure 2: *Architecture of the multilingual phoneme classifier using phonological features.*

mapping[1].

Once the phonological detector is trained, the phonological posteriors gathered from the detectors can be viewed as an indication that a specific phone has been articulated. In this work, the log posteriors of every phonological class are concatenated together and fed into the phoneme classifier, which is realized using a DNN. The outputs of the DNN are the monophone targets of the same IPA-based phoneme set used in multilingual CTC training. For each unseen phoneme in a target language, the phoneme classifier will be utilized to find the most probable mappings in the multilingual phoneme set.

3.3. Parameter Initialization Using Multilingual Phoneme Posterior

When extending the multilingual CTC network to a new language, a better initialization of the parameters connecting to those unseen phonemes can be estimated using the phonological attribute-based phoneme classifier described above. For an unseen phoneme s, the corresponding phonological attributes can be obtained from prior knowledge. Inputting the phonological attributes to the phoneme classifier produces multilingual phoneme posterior $\mathbf{P}(s) = [p_1(s), p_2(s), ..., p_N(s)]$, where N denotes the size of the multilingual phoneme set. The posterior $\mathbf{P}(s)$ can be interpreted as how close the new phoneme is to those seen multilingual phonemes. In the extended output layer, the weights $\mathbf{w}_s$ and the bias b_s of the unseen phoneme s can be initialized either by taking a weighted average of the

[1] http://publications.idiap.ch/downloads/reports/2018/Tong_Idiap-Com-02-2018.pdf

Table 1: *Statistics of the dataset of each language used in this work: the amounts of speech data are in hours.*

	Language	Dataset	Train	Dev	Test
Multilingual Data	EN	WSJ	81h	1.1h	0.7h
	FR	BREF/GP	120h	10.3h	8.8h
	GE	BCN	136h	1.1h	5.7h
	Total Amount		337h		
Target Language	PO	GP	21h	1.6h	1.8h

parameters of all the seen multilingual phonemes,

$$\mathbf{w}_s = \sum_{i=1}^{N} p_i(s)\mathbf{w}_i, b_s = \sum_{i=1}^{N} p_i(s) b_i \quad (1)$$

where $\mathbf{w}_i$ and b_i represent the weight and the bias of the i^{th} phoneme respectively, or by copying the weight and bias of the multilingual phoneme that has the maximum posterior.

$$\mathbf{w}_s = \mathbf{w}_m, b_s = b_m, m = \underset{i}{\operatorname{argmax}}(p_i(s)) \quad (2)$$

4. Experiments

4.1. Experimental Database

The multilingual seed model was trained on English (EN), French (FR), and German (GE). The English data was obtained from the Wall Street Journal (WSJ) corpus [22]. Data preparation gave us 81 hours of transcribed speech. The French data was extracted from the BREF [23] and GlobalPhone (GP) corpora [24], which consist of 120 hours of data. From the German Broadcast News (BCN) corpus [25], we used 136 hours of data for training. In total, 337 hours of multilingual data was used for multilingual CTC training. All the training data is quite clean read speech from similar acoustic conditions. In cross-lingual adaptation experiments, GlobalPhone Portuguese (PO) was considered as the target low-resourced language, which has only 21 hours data. The detailed statistics for each of the languages is shown in Table 1. The development sets were used to tune the hyper parameters for training.

4.2. Setup

We used 40-dimensional log-mel filterbank coefficients as acoustic features together with their first and second-order derivatives, derived from 25 ms frames with a 10 ms frame shift. The features were normalized via mean subtraction and variance normalization on a speaker basis. All the monolingual phones were mapped to IPA symbols and we merged the phonemes from EN, FR and GE to create the universal phone set for multilingual training. The multilingual CTC model has 4 layers of Bidirectional Long Short-Term Memory (BLSTM), with 320 cells in each layer and direction. All the weights in the models were randomly initialized and were trained using stochastic gradient descent with momentum. A learning rate of 0.00004 was used and early stopping on the validation set was applied to select the best model. The dropout rate was set to 0.2.

Once the multilingual model was trained, it was used as seed model for cross-lingual adaptation to Portuguese. A similar training strategy was applied. For decoding, a weighted finite-state transducer (WFST) decoding graph was built using a language-specific lexicon and language model. The trigram language models that we used are publicly available[2]. All the DNN/HMMs compared in this work have 6 hidden layers, each consisting of 1024 units. Thus, it contains slightly more parameters (8.8 vs 8.5 million) than the CTC models. All CTC models were trained using the EESEN implementation [26] and DNN/HMM systems were built using the Kaldi [27].

[2] http://www.csl.uni-bremen.de/GlobalPhone/

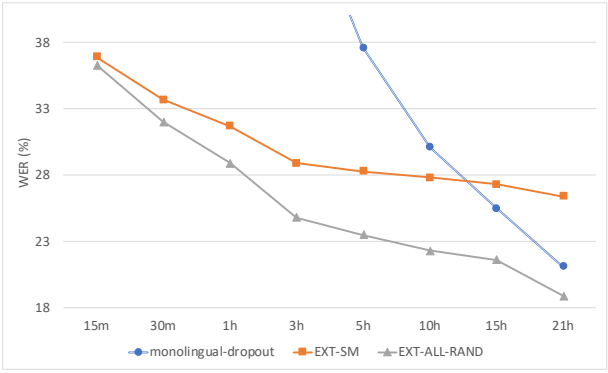

Figure 3: *WERs (%) of different approaches in cross-lingual adaptation. The WERs of monolingual CTC models on less than 5 hours data are above 50% and exceed the graph region.*

4.3. Results

4.3.1. Updating Whole Network vs. Updating Output Layer

In previous work [11], we showed that updating the whole network performs better than only updating the output layer and extending the output layer further improves the performance, as described in Figure 1. However, in the present experiment, we are interested in even smaller data sizes. We hypothesize that updating only the output layer might achieve better performance on more limited data. Therefore, we revisited the comparison between updating the whole network and updating only the output layer after extending the multilingual output layer to Portuguese and also did the comparison on less data (15 min, 30 min). The parameters connecting to unseen phonemes were randomly initialized. Since dropout has been proved to be effective in CTC-based cross-lingual adaptation, it was also applied in this work.

As shown in Figure 3, updating the whole network (EXT-ALL-RAND) consistently outperforms only updating the output layer (EXT-SM) even on 15-30 minutes adaptation data. It further confirms the previous observation. Therefore, all the parameters in the networks were updated with dropout during cross-lingual adaptation in the remaining experiments.

4.3.2. Phonological Attribute Detector and Phoneme Classifier

The same multilingual data, EN, FR and GE was used to train the phonological attribute detector and the phoneme classifier. The phonological attribute detector is a 4 layer DNN, with 1024 hidden units in each layer. The same log-mel filterbank coefficients but with 5 frames context on each side were used as input features. The detector produces greater than 92.2% frame-level attribute detection accuracies for all phonological attributes used in this work and an overall 96.2% accuracy. Because of the limited space, we do not list the detection accuracy for all the attributes.

The input of the phoneme classifier is the concatenated log phonological posteriors with 5 frames context on each side. The DNN has 6 layers, each consisting of 1024 units. The output targets are multilingual IPA monophones based on EN, FR and

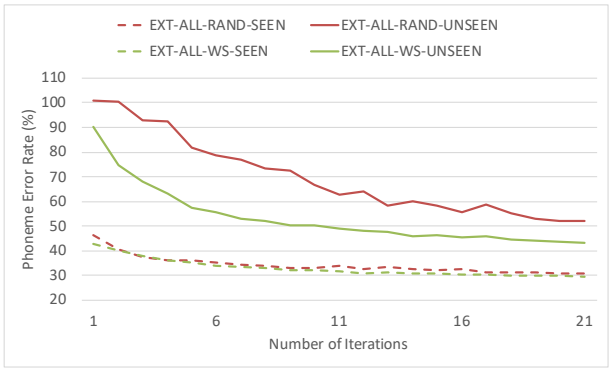

Figure 4: *PERs (%) with respect to overlapped phonemes (SEEN) and new phonemes (UNSEEN) on PO development set. The adaptation was performed on 30 minutes data.*

Table 2: *WERs (%) of cross-lingual adaptation with different initialization. WS denotes weighted summation of the multilingual weights in initialization and MAX represents taking the weights of the most probable mapped phonemes.*

	15m	30m	1h	5h	10h	15h	21h
EXT-ALL-RAND	36.9	32.0	28.9	23.5	22.3	21.6	18.7
EXT-ALL-WS	33.7	29.6	27.7	23.5	22.0	21.2	18.5
EXT-ALL-MAX	34.3	29.7	27.9	23.5	22.2	21.4	18.9

GE, as described above. The test sets from the 3 languages were merged together to test the phoneme classification accuracy. The overall accuracy is 86.4%, which means it is a reliable phoneme predictor using phonological information.

4.3.3. Posterior-based Parameter Initialization

There are 19 phonemes in Portuguese that never appear in the experimental multilingual IPA phoneme set. Phonological attributes can be derived for each of the unseen phonemes based on prior knowledge. Phoneme posteriors were obtained by inputting the phonological attributes. Both parameter initialization approaches were tested.

As shown in Table 2, both posterior-based initialization approaches achieve better performance with less than 3 hours adaptation data. The improvement becomes smaller and smaller with the increase of the adaptation data. As an example, we analyzed the phoneme error rate (PER) with respect to overlapped phonemes and new, unseen, phonemes separately on the development set during CTC training. As plotted in Figure 4. it shows that training from posterior-based initialization keeps the same performance on seen phonemes and yields much better PER on unseen phonemes. When adaptation data is limited, the model initialized using phonological information can quickly catch up on new phonemes.

The two initialization approaches perform almost the same. The phoneme posterior given by the phoneme classifier for each unseen phoneme is quite high, as listed in Table 3. This explains why there is little difference.

4.3.4. Compare CTC-based and DNN/HMM-based Adaptation

We also compared our proposed CTC-based adaptation with DNN/HMM-based adaptation approaches. In the DNN/HMM-based adaptation, the multilingual DNN trained on the same multilingual data was used as the seed model. We then replaced the multilingual output layer with Portuguese targets.

Table 3: *The most probable mappings of the 19 unseen Portuguese phonemes. The numbers in parentheses are the corresponding posteriors. Phonemes are represented in X-SAMPA.*

a"	6 ~	6 ~ "	d_j	e"	e ~ "	i"
a(0.64)	6(0.96)	6(0.96)	d(0.98)	e(0.88)	e ~(0.95)	i(0.96)
i ~	i ~ "	L	o"	o ~ "	r	t_j
i(0.93)	i(0.93)	j(0.93)	o(0.88)	o ~(0.99)	h(0.66)	t(0.98)
u"	u ~	u ~ "	l =	l =~		
u(0.97)	u(0.96)	u(0.96)	l(0.95)	n(0.7)		

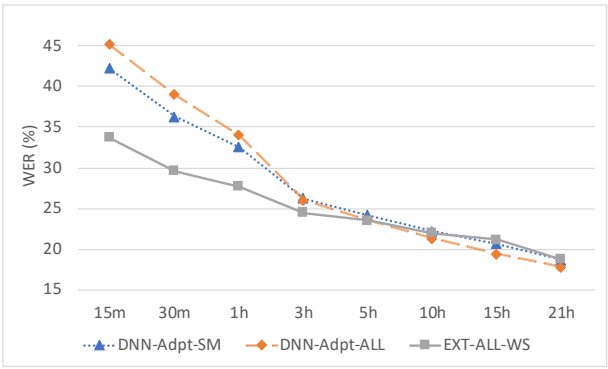

Figure 5: *Comparison between CTC-based and DNN/HMM-based cross-lingual adaptation in WER(%). DNN-Adpt-SM denotes only updating the output layer. DNN-Adpt-ALL represents updating the whole network.*

The Portuguese context-dependent states and alignments were obtained from GMM/HMM systems trained on the corresponding amount of adaptation data. The adaptation was performed by either updating the whole network or only updating the output layer. Dropout was not applied for DNN since performance degradation was observed with dropout in our experiments.

It is clear from Figure 5 that the proposed CTC-based cross-lingual adaptations significantly outperform the DNN/HMM-based models on limited adaptation data (less than 3 hours). CTC-based models retain all the information learned in multilingual training. by contrast, DNN/HMM-based adaptation only keeps the knowledge in hidden layers. This difference makes CTC-based models highly competitive when only limited data is available.

5. Conclusions

It was demonstrated that updating the whole network outperforms only updating the output layer in CTC-based cross-lingual adaptation. When data is extremely limited, leveraging human knowledge and phonological information to initialize the model parameters can make the model converge faster and better. The proposed initialization approach was shown to yield better performance than conventional DNN/HMM-based cross-lingual adaptation on limited data, potentially making the CTC model a competitive alternative in fast language adaptation of an ASR system.

6. Acknowledgement

This work has been conducted with the support of the European Community H2020 Research and Innovation Action-funding, under Scalable Understanding of Multilingual Media (SUMMA) project n. 688139.

7. References

[1] S. Thomas, S. Ganapathy, and H. Hermansky, "Multilingual MLP features for low-resource LVCSR systems," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

[2] K. Knill, M. J. Gales, S. P. Rath, P. C. Woodland, C. Zhang, and S.-X. Zhang, "Investigation of multilingual deep neural networks for spoken term detection," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2013.

[3] F. Grézl, M. Karafiát, and K. Veselỳ, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2014.

[4] J.-T. Huang, J. Li, D. Yu, L. Deng, and Y. Gong, "Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[5] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[6] H. A. Bourlard and N. Morgan, *Connectionist speech recognition: a hybrid approach*, 2012.

[7] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, 2012.

[8] 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.

[9] H. Sak, A. Senior, K. Rao, O. Irsoy, A. Graves, F. Beaufays, and J. Schalkwyk, "Learning acoustic frame labeling for speech recognition with recurrent neural networks," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2015.

[10] Y. Miao, M. Gowayyed, X. Na, T. Ko, F. Metze, and A. Waibel, "An empirical exploration of CTC acoustic models," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016.

[11] S. Tong, P. N. Garner, and H. Bourlard, "Multilingual training and cross-lingual adaptation on CTC-based acoustic model," *arXiv preprint arXiv:1711.10025*, 2017.

[12] S. Stuker, T. Schultz, F. Metze, and A. Waibel, "Multilingual articulatory features," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2003.

[13] S. M. Siniscalchi, D.-C. Lyu, T. Svendsen, and C.-H. Lee, "Experiments on cross-language attribute detection and phone recognition with minimal target-specific training data," *IEEE Transactions on Acoustics, Speech and Signal Processing*, 2012.

[14] S. M. Siniscalchi, D. Yu, L. Deng, and C.-H. Lee, "Exploiting deep neural networks for detection-based speech recognition," *Neurocomputing*, 2013.

[15] S. Kim and M. L. Seltzer, "Towards language-universal end-to-end speech recognition," *arXiv preprint arXiv:1711.02207*, 2017.

[16] 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," *arXiv preprint arXiv:1711.01694*, 2017.

[17] D. Imseng, H. Bourlard, J. Dines, P. N. Garner, and M. Magimai.-Doss, "Improving non-native ASR through stochastic multilingual phoneme space transformations," in *Proceedings of Interspeech*, Florence, Italy, August 2011.

[18] N. Srivastava, G. E. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: a simple way to prevent neural networks from overfitting." *Journal of machine learning research*, 2014.

[19] J. Billa, "Improving LSTM-CTC based ASR performance in domains with limited training data," *arXiv preprint arXiv:1707.00722*, 2017.

[20] S. Dalmia, R. Sanabria, F. Metze, and A. W. Black, "Sequence-based multi-lingual low resource speech recognition," *arXiv preprint arXiv:1802.07420*, 2018.

[21] R. Rasipuram and M. Magimai-Doss, "Improving articulatory feature and phoneme recognition using multitask learning," in *International Conference on Artificial Neural Networks*, 2011.

[22] D. B. Paul and J. M. Baker, "The design for the Wall Street Journal-based CSR corpus," in *Proceedings of the workshop on Speech and Natural Language*, 1992.

[23] L. F. Lamel, J.-L. Gauvain, M. Eskénazi *et al.*, "BREF, a large vocabulary spoken corpus for french1," 1991.

[24] T. Schultz, N. T. Vu, and T. Schlippe, "GlobalPhone: A multilingual text & speech database in 20 languages," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[25] F. Weninger, B. Schuller, F. Eyben, M. Wöllmer, and G. Rigoll, "A broadcast news corpus for evaluation and tuning of German LVCSR systems," *arXiv preprint arXiv:1412.4616*, 2014.

[26] Y. Miao, M. Gowayyed, and F. Metze, "EESEN: End-to-end speech recognition using deep RNN models and WFST-based decoding," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2015.

[27] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlicek, Y. Qian, P. Schwarz *et al.*, "The Kaldi speech recognition toolkit," in *Proceedings of the IEEE Workshop on Automatic Speech Recognition and Understanding*, 2011.