SEQUENCE-BASED MULTI-LINGUAL LOW RESOURCE SPEECH RECOGNITION

Siddharth Dalmia, Ramon Sanabria, Florian Metze and Alan W. Black

Language Technologies Institute, Carnegie Mellon University; Pittsburgh, PA; U.S.A.
{sdalmia|ramons|fmetze|awb}@cs.cmu.edu

ABSTRACT

Techniques for multi-lingual and cross-lingual speech recognition can help in low resource scenarios, to bootstrap systems and enable analysis of new languages and domains. End-to-end approaches, in particular sequence-based techniques, are attractive because of their simplicity and elegance. While it is possible to integrate traditional multi-lingual bottleneck feature extractors as front-ends, we show that end-to-end multi-lingual training of sequence models is effective on context independent models trained using Connectionist Temporal Classification (CTC) loss. We show that our model improves performance on Babel languages by over 6% absolute in terms of word/phoneme error rate when compared to mono-lingual systems built in the same setting for these languages. We also show that the trained model can be adapted cross-lingually to an unseen language using just 25% of the target data. We show that training on multiple languages is important for very low resource cross-lingual target scenarios, but not for multi-lingual testing scenarios. Here, it appears beneficial to include large well prepared datasets.

Index Terms— multi-lingual speech recognition, cross-lingual adaptation, connectionist temporal classification, feature representation learning

1. INTRODUCTION

State-of-the-art speech recognition systems with human-like performance [1, 2] are trained on hundreds of hours of well-annotated speech. Since annotation is an expensive and time-consuming task, similar performance is typically unattainable on low resource languages. Multi-lingual or cross-lingual techniques allow transfer of models or features from well-trained scenarios to those where large amounts of training data may not be available, cannot be transcribed, or are otherwise hard to come by [3, 4].

The standard approach is to train a context dependent Hidden Markov Model based Deep Neural Network acoustic model with a "bottleneck" layer using a frame based criterion on a large multi-lingual corpus [5, 6, 7]. The network up to the bottleneck layer can be used as a language-independent feature extractor while adapting to a new language. Generating such a model requires the preparation of frame level segmentation in each language, which is usually achieved by training separate mono-lingual systems first. This is a cumbersome multi-step process. Moreover, if the speaking style, acoustic quality, or linguistic properties of the recordings are very different across a set of languages, the segmentations may be inconsistent across languages and thus sub-optimal for generating features in a new language.

On the other hand, end-to-end training approaches which directly model context independent phones are elegant, and greatly facilitate speech recognition training. Most do not require an explicit alignment of transcriptions with the training data, and there are typically fewer hyper-parameters to tune. We show that sequence training in multi-lingual settings can create feature extractors, which can directly be ported to new languages using a linear transformation (on very limited data), or re-trained on more data, opening a door to end-to-end language universal speech recognition.

2. RELATED WORK AND BABEL DATASET

Some of the early works in multi-lingual and cross-lingual speech recognition involved the use of language independent features like articulatory features [8] to train HMM based systems. Authors in [9] used subspace Gaussian mixture model to map phonemes of different languages together. Authors in [10] introduce the use of a shared phone set to build HMM based language independent acoustic models and show the adaptation of pre-existing models towards a new language.

With the on-set of deep learning the focus of the models shifted to learning features across languages which can be mapped to the same space [3, 11]. Authors in [12] looked at unsupervised pretraining on different languages for a cross lingual recognition. The dominant architecture for multi-lingual or cross-lingual speech recognition has

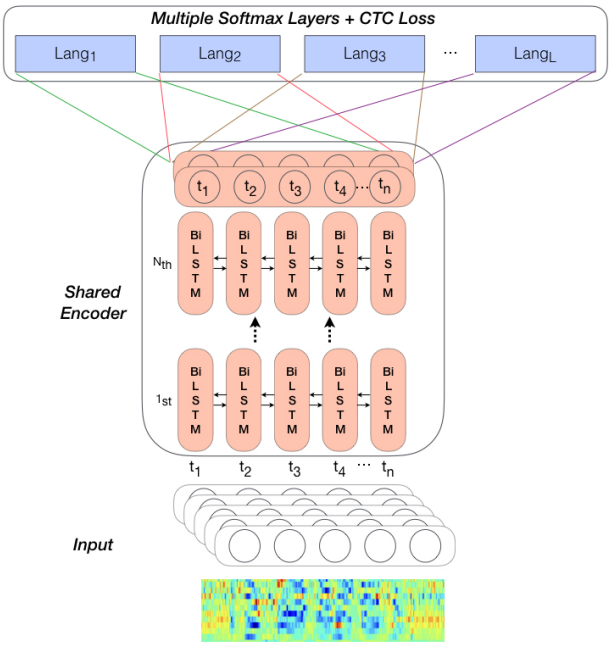


Fig. 1. Multi-lingual CTC model following the "shared hidden layer" approach for LSTM layers.

been the so-called "shared hidden layer" model, in which data is passed through a series of shared feed-forward layers, before being separated into multiple language-specific softmax layers, which are trained using cross-entropy [13, 5, 14]. This architecture can also be used as a "bottleneck" feature extractor, from which "language independent" features are extracted, on top of which a target-language acoustic model can be built. Authors in [15] showed that these multilingual models can be adapted to the specific language to improve performance further. The work by [5, 16] presented bottleneck features for multi-lingual systems where they showed feature porting is possible and gave competitive results when compared to systems with mono-lingual features. Other approaches [17, 18] constructed a shared language independent phone set, which could then also be adapted to the target language. Our proposed model is inspired by the former approach which tries to learn latent features by sharing hidden layers across languages.

Connectionist Temporal Classification (CTC, [19]) lends itself to low-resource multi-lingual experiments, because systems built on CTC tend to be significantly easier to train than those that have been trained using hidden Markov models [20, 21]. [22] shows that multi-lingual CTC systems with shared phones can improve performance in a limited data setting. As per our knowledge there has not been any prior work that have looked into learning "bottleneck" like features for a CTC based model and seen how it performs multilingually and cross-lingually with adaptation.

For this paper we use several languages from IARPA's Babel[1] project to test our model. These are mostly telephony (8kHz) conversational speech data in a low resource language. These were accompanied by a lexicon and dictionary in Extended Speech Assessment Methods Phonetic Alphabet (X-SAMPA) format. Table 1 summarizes the amount of training data in hours along with the number of phonemes (including the CTC blank symbol) present for the languages we used in our experiments on the "Full Language Pack" (FLP) condition.

Table 1. Overview of the FLP Babel Corpora used in this work.

Subset	Language	# Phones + $\emptyset$	Training Data
MLing	Turkish	50	79 hrs
	Haitian	40	67 hrs
	Kazakh	70	39 hrs
	Mongolian	61	46 hrs
Bab300	Amharic	67	43 hrs
	Tamil	41	69 hrs
	Tagalog	48	85 hrs
	Pashto	54	78 hrs
For testing	Kurmanji	45	42 hrs
	Swahili	40	44 hrs

3. MULTI-LINGUAL CTC MODEL

A model trained with CTC loss is a sequence based model which automatically learns alignment between input and output by introducing an additional label called the blank symbol ($\emptyset$), which corresponds to 'no output' prediction. Given a sequence of acoustic features $\mathbf{X} = (\mathbf{x_1}, \ldots, \mathbf{x_n})$ with the label sequence $\mathbf{z} = (\mathbf{z_1}, \ldots, \mathbf{z_u})$, the model tries to maximize the likelihood of all possible CTC paths $\mathbf{p} = (\mathbf{p_1}, \ldots, \mathbf{p_n})$ which lead to the correct label sequence $\mathbf{z}$ after reduction. A reduced CTC path is obtained by grouping the duplicates and removing the $\emptyset$ (e.g. $\mathcal{B}(AA\emptyset AABBC) = AABC$).

$$P(\mathbf{z}|\mathbf{X}) = \sum_{\mathbf{p} \in \texttt{CTC\_Path}(\mathbf{z})} P(\mathbf{p}|\mathbf{X})$$

Like in [20] we use this loss along with stacked Bidirectional LSTM layers to encode the acoustic information and make frame-wise predictions.

In our CTC multi-lingual model, we share the bidirectional LSTM encoding layer till the final layer and project the learned embedding layer to the phones of the respective target languages. The intuition behind this model is that training on more than one language will help in better regularization of weights and learning a better representation of features, as it will be trained on more data. We hypothesize that the final phoneme discrimination can be learned in a linear projection of the last layer. Figure 1 shows the schematic diagram of our multi-lingual model. Mathematically this can be written as,

$$\mathbf{X} = \{\mathbf{X}_{L1} \cup \mathbf{X}_{L2} \cup \mathbf{X}_{L3} \ldots \mathbf{X}_{Ln}\} \quad \mathbf{X}_{Li} = (x^1_{Li}, \ldots, x^n_{Li})$$
$$\mathbf{e} = \texttt{Encoder}_{BiLSTM}(X) \quad \mathbf{e} \in \mathbb{R}^{n \times 2 * h_{dim}}$$

$$P(\mathbf{p}|\mathbf{X}) = \begin{cases} \texttt{softmax}(\mathbf{W}_{L1}\mathbf{e} + \mathbf{b}_{L1}) & \text{if } \mathbf{X} \in \mathbf{X}_{L1} \\ \texttt{softmax}(\mathbf{W}_{L2}\mathbf{e} + \mathbf{b}_{L2}) & \text{if } \mathbf{X} \in \mathbf{X}_{L2} \\ \ldots & \\ \texttt{softmax}(\mathbf{W}_{Ln}\mathbf{e} + \mathbf{b}_{Ln}) & \text{if } \mathbf{X} \in \mathbf{X}_{Ln} \end{cases}$$

Unlike [5], we do not have any bottleneck layer, and the whole model is sequence trained based on CTC loss. Note that here we recognize a sequence of phonemes which is a much harder problem. Traditional HMM/DNN systems perform frame-wise recognition of individual phonemes, usually relying on alignments that have been generated by mono-lingual models. This can be considered a much simpler task than the recognition of a phone sequence.

4. EXPERIMENTS AND OBSERVATIONS

4.1. Multi-lingual CTC model

To align with project goals, we chose to perform experiments on a set of four languages which are the closest/ have maximum phone overlap with Kurmanji – Kazakh, Turkish, Mongolian and Haitian. We used a 6-layer bidirectional LSTM network with 360 cells in each direction, which performed best on average across the majority of Babel languages in a systematic search experiment. Table 2 shows the results. For consistency, we used absolutely identical settings across all languages, and did not perform any language-specific tuning, other than choosing the lowest perplexity language model between 3-gram and 4-gram models for WFST-based decoding. Techniques such as blank scaling and applying a softmax temperature can often improve results significantly, but we did not apply any of them here for consistency.

In our multi-lingual experiments, we use the same 6-layer Bi-LSTM network with 360 cells (per direction) in each layer as our shared encoded representation[2]. Again, this setup performed best on average on a larger set of languages. Multi-lingual training on

[1] This work used releases IARPA-babel105b-v0.4, IARPA-babel201b-v0.2b, IARPA-babel401b-v2.0b, IARPA-babel302b-v1.0a (these 4 languages will be called the "MLing" set), and IARPA-babel106b-v0.2g, IARPA-babel307b-v1.0b, IARPA-babel204b-v1.1b, IARPA-babel104b-v0.4bY (these 4 languages will be called the "BAB300" set), and IARPA-babel202b-v1.0d and IARPA-babel205b-v1.0 for testing.

[2] The code to train the multi-lingual model will be released as part of EESEN [20].

Table 2. Word (% WER) and phoneme error rate (% PER) for each of the test languages, on the Babel conversational development test sets.

Model	Kazakh		Turkish		Haitian		Mongolian	
	WER	PER	WER	PER	WER	PER	WER	PER
Mono-lingual	55.9	40.9	53.1	36.2	49.0	36.9	58.2	45.2
Multi-lingual (MLing)	53.2	36.5	52.8	34.4	47.8	34.9	55.9	41.1
MLing & FineTuning (FT)	50.6	35.1	49.0	32.2	46.6	33.2	53.4	39.6
MLing + SWBD	52.3	36.6	51.3	33.0	45.8	33.9	54.5	40.2
MLing + SWBD & FT	**48.2**	**33.5**	**48.7**	**31.9**	**44.3**	**31.9**	**51.5**	**37.8**

the "MLing" set (the four languages shown in Table 2) improves WER by 1.7% (absolute) on average, while keeping the LSTM layers shared across all languages. If we fine-tune the entire model towards each language specifically, performance improves further, by 4.4% on average over the baseline. If we roughly double the amount of training data by adding the Switchboard 300h training set to the "MLing" training data, performance improves yet again, for both the universal (MLing+SBWD) and language-specific (MLing+SWBD & FT) case. Overall, WER and PER improve by about 6% absolute (>10% relative), which is in line with other results reported on comparable tasks discussed in section 2.

As expected, reductions in the error rates tend to be higher for the lower resource languages, like Kazakh and Mongolian.

4.2. Data Selection

Given that adding a seemingly unrelated, but high resource language improved the performance of the model on four low resource languages, we further studied the impact of varying the source(s) of the extra data. Specifically, we replaced the 300 h Switchboard corpus with four more unrelated Babel languages, "BAB300" composed of Tamil, Amharic, Pashto, and Tagalog. The results on the test data are summarized in Table 3. We can see that adding Switchboard data outperforms adding more unrelated Babel languages.

Table 3. Word error rate (% WER) on the test languages when switching the SWBD data with 300 hrs equivalent of Babel.

Model	Kazakh	Turkish	Haitian	Mongolian
MLing + BAB300	57.5	52.0	47.8	56.7
MLing + SWBD	**52.3**	**51.3**	**45.8**	**54.5**

While our main goal here has been the creation of a multi-lingual recognizer, we verified that models that have been trained on a single Babel language plus 300 h of Switchboard do not outperform the fine-tuned MLing+SWBD system, while there is no clear pattern on other languages. This indicates that it is generally beneficial to train (sequence-based) multi-lingual systems on closely related languages, and/or on large amounts of well-prepared but unrelated mono-lingual data, but that adding a large number of languages may in fact prevent the model from training well.

4.3. Representation Learning

In order to study to what extent the CTC sequence models have learned useful bottleneck like discriminatory audio features that are independent of the input language, we attempt to port a model to an unseen language. We aim to use the trained model as a language-independent feature extractor that can linearly separate any language into a phoneme sequence. To do this, we replace the softmax layer (or "layers" in the multi-lingual case) of a "donor" CTC model with a single softmax, which we then train with varying amounts of data from the target language, Kurmanji in our case. Figure 2 shows how different "donor" models behave in this situation. In the cross-lingual case, it becomes beneficial to train the LSTM layers with as many different languages as possible ("MLing+BAB300" outperforms "MLing+SWBD" and "MLing"), while a single related language (Turkish) outperforms adaptation on a larger amount of data from an unrelated language (SWBD). There is a large gap between mono-lingual systems and multi-lingual systems. Improvements become smaller once training is performed on 4 h (10%) of data or more, but even then the re-estimation of the softmax layer (with ca. 32k parameters) benefits from more data.

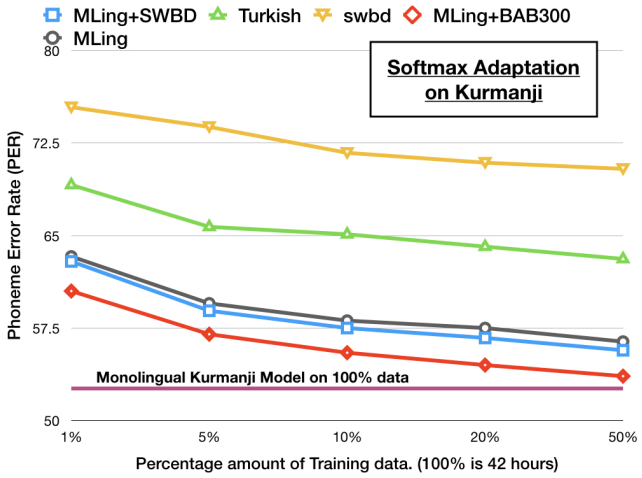


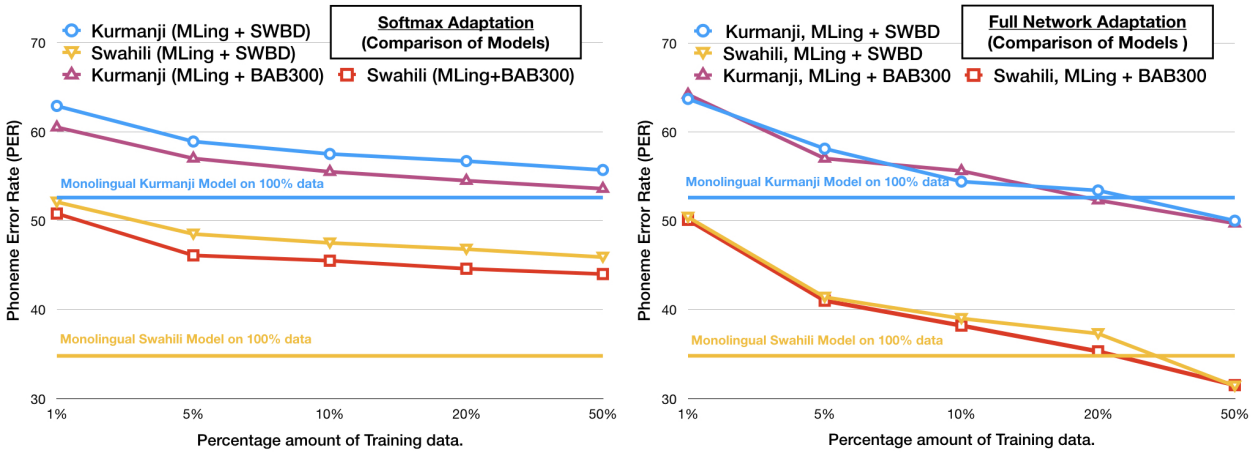
Fig. 2. Cross-lingual training of CTC softmax layer only on top of different "donor" models.

It thus seems that multi-lingual systems do indeed learn a portable, language independent representation, which is useful when porting to a new language, while the sheer amount of data is less beneficial.

4.4. Cross-lingual Explorations

Figure 4 shows that for both related and unrelated languages, a multi-lingual system surpasses the mono-lingual baseline once about 25% of the original data has been seen. The behavior of retraining ("full network adaptation") seems independent of the original trained languages.

To further investigate how multi-lingual models can be used in cross-lingual settings, and with varying amounts of training data, we



(a) Adaptation of softmax layer only for Kurmanji and Swahili targets. Kurmanji performs well, because the language is similar to some training languages.

(b) Adaptation of entire network (re-training) to target languages. This outperforms softmax adaptation (on the left) as soon as 2-4 h of data become available.

Fig. 3. Cross-lingual training of Kurmanji and Swahili systems.

compare "softmax" adaptation and full network adaptation (retraining) on Kurmanji and Swahili, two languages which we did not see in training. We use the (MLing + SWBD) and (MLing + BAB300) "donor" models. Figure 3 shows that for small amounts of adaptation data, and a target language that is related to the pre-trained languages (Kurmanji), "softmax adaptation" is competitive, and an initialization with many languages is beneficial.

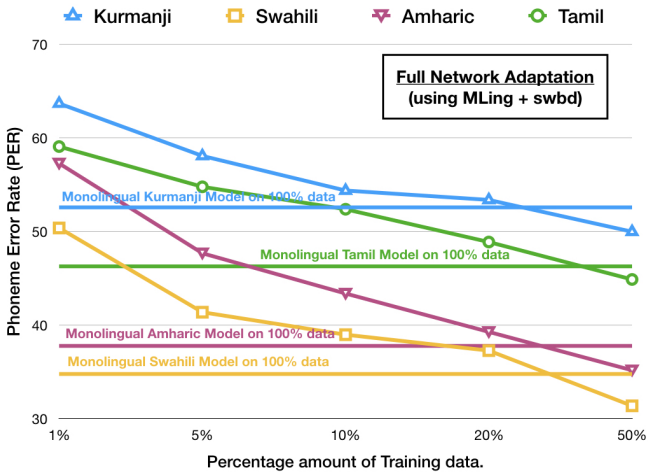


Fig. 4. PER on different amounts of cross-lingual data using a full network end-to-end adaptation (retraining).

When the entire network can be retrained ("full network adaptation", shown on the right side of Figure 3), there is very little difference between the "donor" systems' performance.

5. CONCLUSION

In this paper, we demonstrate that it is possible to train multi-lingual and cross-lingual acoustic models directly on phone sequences, rather than frame-level state labels. Unlike multi-lingual bottleneck features, these CTC models do not require the generation of state alignments, which facilitates their use.

In multi-lingual settings, it seems beneficial to train on related languages only, or on large amounts of clean data; there is no benefit simply from training on many languages. It is thus possible to combine e.g. Switchboard and Babel data.

In very low resource cross-lingual scenarios, it is possible to adapt a model to a previously unseen language by re-training the softmax layer only. CTC models can learn a language independent representation at the input to the softmax layer. We find that training the models trained on related languages help, as does training on many languages, rather than large amounts of data. As more and more data is available, and the whole network can be retrained, and the effect of the choice of language for the multi-lingual training disappears.

As future work, we are investigating on decoding the CTC output using a phoneme based neural language models trained on non-parallel text, thereby facilitating us to do zero-resource speech recognition.

6. ACKNOWLEDGEMENTS

This project was sponsored by the Defense Advanced Research Projects Agency (DARPA) Information Innovation Office (I2O), program: Low Resource Languages for Emergent Incidents (LORELEI), issued by DARPA/I2O under Contract No. HR0011-15-C-0114.

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research. This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number OCI-1053575. Specifically, it used the Bridges system, which is supported by NSF award number ACI-1445606, at the Pittsburgh Supercomputing Center (PSC).

We are grateful to Anant Subramanian and Soumya Wadhwa for their feedback on the presentation of this paper.

7. REFERENCES

[1] G. Saon, G. Kurata, T. Sercu, K. Audhkhasi, S. Thomas, D. Dimitriadis, X. Cui, B. Ramabhadran, M. Picheny, L.-L. Lim, et al., "English conversational telephone speech recognition by humans and machines," *arXiv preprint arXiv:1703.02136*, 2017.

[2] W. Xiong, J. Droppo, X. Huang, F. Seide, M. Seltzer, A. Stolcke, D. Yu, and G. Zweig, "Achieving human parity in conversational speech recognition," *arXiv preprint arXiv:1610.05256*, 2016.

[3] A. Stolcke, F. Grezl, M.-Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Acoustics, Speech and Signal Processing, 2006. ICASSP 2006 Proceedings. 2006 IEEE International Conference on*. IEEE, 2006, vol. 1, pp. I–I.

[4] F. Grézl, E. Egorova, and M. Karafiát, "Study of large data resources for multilingual training and system porting," *Procedia Computer Science*, vol. 81, pp. 15–22, 2016.

[5] K. Veselỳ, M. Karafiát, F. Grézl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 336–341.

[6] 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 *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 138–143.

[7] N. T. Vu, F. Metze, and T. Schultz, "Multilingual bottle-neck features and its application for under-resourced languages," in *Proc. 3rd Workshop on Spoken Language Technologies for Under-resourced Languages*, Cape Town; S. Africa, May 2012, MICA.

[8] S. Stuker, F. Metze, T. Schultz, and A. Waibel, "Integrating multilingual articulatory features into speech recognition," in *Eighth European Conference on Speech Communication and Technology*, 2003.

[9] L. Burget, P. Schwarz, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. Goel, M. Karafiát, D. Povey, et al., "Multilingual acoustic modeling for speech recognition based on subspace Gaussian mixture models," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4334–4337.

[10] T. Schultz and A. Waibel, "Language-independent and language-adaptive acoustic modeling for speech recognition," *Speech Communication*, vol. 35, no. 1, pp. 31–51, 2001.

[11] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 7319–7323.

[12] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in dnn-based lvcsr," in *Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 246–251.

[13] S. Scanzio, P. Laface, L. Fissore, R. Gemello, and F. Mana, "On the use of a multilingual neural network front-end," ISCA, 2008.

[14] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual acoustic models using distributed deep neural networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 8619–8623.

[15] F. Grézl, M. Karafiát, and K. Vesely, "Adaptation of multilingual stacked bottle-neck neural network structure for new language," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7654–7658.

[16] F. Grézl, M. Karafiat, and M. Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Automatic Speech Recognition and Understanding (ASRU), 2011 IEEE Workshop on*. IEEE, 2011, pp. 359–364.

[17] S. Tong, P. N. Garner, and H. Bourlard, "An Investigation of Deep Neural Networks for Multilingual Speech Recognition Training and Adaptation," in *Proc. of Interspeech*, 2017, number EPFL-CONF-229214.

[18] N. T. Vu, D. Imseng, D. Povey, P. Motlicek, T. Schultz, and H. Bourlard, "Multilingual deep neural network based acoustic modeling for rapid language adaptation," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 7639–7643.

[19] 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*. ACM, 2006, pp. 369–376.

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

[21] Y. Miao, M. Gowayyed, X. Na, T. Ko, F. Metze, and A. Waibel, "An empirical exploration of CTC acoustic models," in *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE, 2016, pp. 2623–2627.

[22] M. Müller, S. Stüker, and A. Waibel, "Language Adaptive Multilingual CTC Speech Recognition," in *Proc. SPECOM*, Hatfield, UK, 2017.