Language Adaptive Multilingual CTC Speech Recognition

Markus Müller[1], Sebastian Stüker[1], and Alex Waibel[1,2]

[1] Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany,
m.mueller@kit.edu,
http://isl.anthropomatik.kit.edu
[2] Language Technology Institute, Carnegie Mellon University, Pittsburgh PA, USA

Abstract. Recently, it has been demonstrated that speech recognition systems are able to achieve human parity. While much research is done for resource-rich languages like English, there exists a long tail of languages for which no speech recognition systems do yet exist. The major obstacle in building systems for new languages is the lack of available resources. In the past, several methods have been proposed to build systems in low-resource conditions by using data from additional source languages during training. While it has been shown that DNN/HMM hybrid setups trained in low-resource conditions benefit from additional data, we are proposing a similar technique using sequence based neural network acoustic models with Connectionist Temporal Classification (CTC) loss function. We demonstrate that setups with multilingual phone sets benefit from the addition of Language Feature Vectors (LFVs).

Keywords: speech recognition, low-resource, multilingual training, connectionist temporal classification

1 Introduction

In recent years, the use of artificial neural networks (ANNs) has lead to dramatic improvements in the field of automatic speech recognition (ASR), lately achieving human parity [42, 27]. ANNs are being used as part of the pre-processing pipeline, e.g., for dimensionality reduction [13], or as part of the acoustic model in DNN/HMM hybrid systems. Latest developments include sequence based ANN based setups with Connectionist Temporal Classification (CTC) [9] loss function. Such systems do not require certain types of resources traditional model do, like time-aligned labels, HMMs and cluster trees. Popular network topologies to use in such setups are bi-directional Long-Short Term Memory (LSTM) networks [14].

While proposed back in 2006, this method has gained popularity quite recently, due to the availability of increased computing power that enabled using large amounts of training data. One of the main advantages of CTC based systems over conventional speech recognition systems is that they are able to

capture temporal dependencies by themselves. While HMM based systems use context-dependent states to mitigate the error made by the Markov assumption (the current state only depends upon the previous state), CTC based systems learn to model context implicitly by the use of Recurrent Neural Networks (RNNs).

But CTC based models are more sensitive to the amount of available training data. This is especially problematic if only a limited amount of data is available during training. In this work, we are proposing a method for adding data from additional source languages. Similar to methods proposed for DNN/HMM based systems, we use data from multiple languages during training. To train our setup truly multilingual, we use a global phone set combining the phone sets of all source languages. In addition, we demonstrate that the recognition performance can be improved by the addition of Language Feature Vectors (LFVs) [23]. By applying this proposed method, multilingual systems outperform monolingual systems trained on the target language only.

This paper is organized as follows: In the next Section, we provide an overview of related work in the field. We describe our approach in Section 3, followed by the description of the experimental setup in Section 4. Section 5 contains the results and we conclude this paper in Section 6, where we also provide an outlook to future work.

2 Related Work

2.1 GMM Based Multi- and Crosslingual Systems

Prior to using neural networks as part of speech recognition systems, the use of GMM/HMM based systems was common. The problem of training systems multi- and crosslingually has been addressed in the past to handle data sparsity [41, 32]. Techniques for adapting the cluster tree were proposed [33], but methods for crosslingual adaptation exist as well [36].

2.2 Multilingual DBNFs

Building DNN-based systems in low resource conditions is challenging, especially because DNNs are a data-driven method with many parameters to be trained. Hence, a large amount of data is required for the model to generalize. Several methods have been proposed to use data from additional source languages. The first step is to pre-train models unsupervised, which is language independent [38]. For fine-tuning, several approaches exist to incorporate multilingual data. One possibility is to share the hidden layers between languages, but use language specific output layers [8, 29, 12, 39]. Instead of having independent output layers, block softmax can also be applied [11]. By partitioning the output layer or using language specific output layers, the systems use separate phone sets for each language instead of a global phone set. In general, training DNNs using data from multiple languages in parallel can be considered as a form of multi-task learning [5, 22].

2.3 Neural Network Adaptation

A common method to adapt neural networks to different speakers is the use of i-Vectors [28] or Bottleneck Speaker Vectors (BSVs) [15]. By using such vectors, speaker adaptive neural networks [21] can be built. These low dimensional vectors encode speaker peculiarities which enable the network to adapt to different speaker characteristics. These methods demonstrate that neural networks benefit from additional input modalities.

Similar to BSVs, we have shown that feature augmentation can also be used to adapt ANNs to different languages when trained multilingually. Providing the language identity information using one-hot encoding leads to improvements [25], but does not provide any language characteristics to the network. Language Feature Vectors (LFVs) [23, 24] have shown to encode such language peculiarities, even if the LFV net was not trained on the target language.

2.4 CTC based systems

While originally proposed in 2006 [9], CTC-based systems are becoming more popular these days. Systems can be trained using either phones or graphemes as labels, or jointly together [6]. Recently, a setup being trained directly on words has been proposed [34]. The notion of multi-task learning can also be applied to CTC-based setups [16, 18, 26].

3 Language Adaptive Multilingual CTC

We aimed at training our setup multilingually, opting for using phones over characters. By merging pronunciation dictionaries from multiple languages, we created a global phone set. While there are many approaches of training CTC systems directly on characters and omitting the pronunciation dictionary, we used phones as targets in this first approach because characters or groups of characters are pronounced very differently between languages, e.g., "*th*" in English or "*sch*" in German. Being language independent, phones are always pronounced in the same way, but eventually with a language specific twang. This might introduce classification errors as the network might have difficulties identifying the correct phone independent of the language. Another issue might have been language dependent phone contexts. While HMM-based systems in general suffer in performance when using a multilingual phone set, special techniques have been proposed to adapt the set of context-dependent states (see Section 2.1). CTC-based systems potentially do not suffer as much from this problem because all phone contexts are learned implicitly by the network. In order to compensate for language dependent peculiarities, we used LFVs which have shown to improve the performance of multilingual HMM-based systems.

We based our setup on Baidu's Deepspeech 2 architecture [4]. The network topology is shown in Figure 1. The input features were first processed by 2 2D convolution layers. Convolutional Neural Networks (CNNs) are based on the

idea of Time-delay Neural Networks (TDNNs) [40]. By applying 2D convolution on the spectrogram, these 2D TDNN layers learn filters to extract features in both the frequency and time dimension. We added LFVs to the output of the convolution layers as input to the bi-directional LSTM [14] layers by appending them to the feature vector. The output layer was a fully connected feed-forward layer with softmax activations. In a series of experiments, we evaluated different hidden layer sizes and different amounts of hidden layers to determine the optimal hyper parameter configuration.

During decoding / testing, we did not apply any advanced techniques like WFST decoding [20] or incorporated an external language model. Instead, we used a naive argmax decoding and computed the label error rate (LER), which is similar to the phone error rate (PER), but also accounts for incorrect word separations.

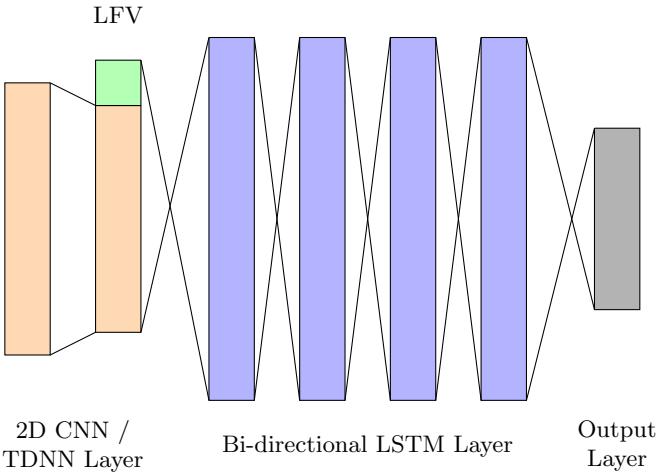

Fig. 1. Network layout, based on Deepspeech2 configuration. LFVs are being added after final convolution layer.

4 Experimental Setup

We based our experiments on the Euronews corpus [10]. It features TV broadcast news recordings from 10 languages. For each language, 70h of data is available, as shown in Table 1. We filtered utterances being shorter than 1s and removed utterances with long phonetic transcripts, because the CUDA implementation supported a maximum label length of 639 symbols[3]. We used only half of the available data per language (approx. 35h) to simulate a resource-constraint task

[3] see: https://github.com/baidu-research/warp-ctc, accessed 2017-04-13

and set aside 10% for testing. We trained systems for both English and German, as well as a system trained jointly on data from both languages.

Table 1. Overview Euronews Corpus

Language	Audio Data	# Recordings
Arabic	72.1h	4,342
English	72.8h	4,511
French	68.1h	4,434
German	73.2h	4,436
Italian	77.2h	4,464
Polish	70.8h	4,576
Portuguese	68.3h	4,456
Russian	72.2h	4,418
Spanish	70.5h	4,231
Turkish	70.4h	4,385
Total	715.6h	44,253

We used MaryTTS [30] to create pronunciations for words contained in the transcriptions. MaryTTS supports multiple languages, with each language having their own set of symbols representing phones. While most of the symbols represent the same phones across languages, we manually mapped symbols which did not match to ensure same phones shared the same symbol. For matching the symbols, we used the definitions of articulatory features embedded in MaryTTS' language definition files. This allowed us to derive a global phone set. Additionally, MaryTTS used special marks to indicate long vowels. As preliminary experiments indicated, the network had difficulties distinguishing between short and long instances of the same vowel. Hence, we discarded marks indicating long vowels. The phone count after and prior to mapping is shown in Table 2. Merging the sound inventory of both languages resulted on a set of 56 phones.

Table 2. Size of different phone sets

Language	Phone Set	Size
English	MaryTTS	42
	Mapped	39
German	MaryTTS	59
	Mapped	48
Combined	Merged	56

To extract acoustic input features, we used the Janus Recognition Toolkit (JRTk) [3], which features the IBIS single-pass decoder [35]. We used our stan-

dard pre-processing pipeline consisting of 40 dimensional log Mel scaled coefficients, as well as 14 dimensional tonal features (FFV [17] and pitch [31]). Adding tonal features even for non-tonal languages has shown improvements [19]. We extracted the features using a window size of 32ms and a frame shift of 10ms. To train the networks, we used PyTorch [1], which provided Python bindings to Torch [7], as well as warp-ctc [2] for computing the CTC loss during network training. The networks were trained using stochastic gradient descent (SGD) with Nesterov momentum [37], a learning rate of 0.0003 and momentum of 0.9. Mini-batch updates with a batch size of 20 and batch normalization were used. Annealing was applied to the learning rate every epoch with a value of 1.1. To prevent gradients from exploding, a max norm constraint of 400 was enforced. During the first epoch, the network was trained with utterances sorted ascending by length.

5 Results

In this section, we first present monolingual results as baseline, followed by the evaluation of different hyper parameter configurations. We then combine data from multiple languages to train a multilingual system and also evaluate adding LFVs to our setup.

5.1 Baseline

As baseline experiment, we trained monolingual systems on English and German using 4 LSTM layers with 400 neurons each. We evaluated the mapping of phones from MaryTTS to actual phone targets of our system. Table 3 shows the results. Using the original phone set from MaryTTS does result in the highest LER, for both English and German.

Table 3. Monolingual results on test set showing the label error rate (LER)

System	Phone Set	LER
English	MaryTTS	20.4%
English	Mapped	**19.0**%
German	MaryTTS	16.0%
German	Mapped	**15.5**%

5.2 Multilingual Experiments

Next, we trained networks multilingually and also evaluated different network hyper parameters. While we kept the configuration of the 2 2D CNN / TDNN layers identical, we varied the parameters of the LSTM layers. For reference, we

also included corresponding results of a monolingual system trained on English. As shown in Table 4, we observed gains from increasing the layer size. But we could not increase the size of the LSTM layers beyond 1,000 neurons per layer because of limitations in GPU memory. Adding an additional layer did not improve the LER.

Table 4. Multilingual results showing the label error rate (LER) for different network configurations

LSTM layer size	**# LSTM layers**	**LER ML**	**LER EN**
350	5	19.6%	–
400	4	20.0%	19.0%
400	5	19.6%	–
600	4	17.3%	–
800	4	16.9%	17.8%
800	5	17.0%	–
1000	**4**	**16.3%**	**17.7%**

5.3 Language Adaptive Networks

Based on the best network configuration (1,000 nodes per layers, 4 LSTM layers), we added LFVs after the CNN / TDNN layers and evaluated the performance of the network for both English and German, as well as multilingually. The results are shown in Table 5. Adding LFVs lowered the LER in all cases. After 7 epochs, the gain over the baseline was bigger on English (8% rel.), compared to German (6% rel.). Training the nets for 70 epochs results in a slight decrease in performance multilingual over monolingual.

Table 5. Multilingual results showing the label error rate (LER)

System	**Monolingual**	**Multilingual**	**LFV**	**LER (7 ep.)**	**LER (70 ep.)**
English	x	–	–	17.7%	13.1%
	–	x	–	18.7%	14.8%
	–	x	x	**16.4**%	**13.5**%
German	x	–	–	14.6%	10.8%
	–	x	–	14.0%	11.8%
	–	x	x	**13.8**%	**11.0**%
Combined	–	x	–	16.3%	12.9%
	–	x	x	**15.7**%	**12.4**%

6 Conclusion

We have presented a method for training CTC based speech recognition systems multilingually. By using LFVs in addition to acoustic input features, we could improve the recognition performance of our multilingual systems. Future work includes the evaluation of additional language combinations and different mixtures of training data. We also intent to use additional adaptation methods like i-Vectors to adapt the networks to different speakers, as well as to further optimize the network architecture and the training process.

References

1. PyTorch. http://pytorch.org, accessed: 2017-04-13
2. warp-ctc. https://github.com/baidu-research/warp-ctc, accessed: 2017-04-13
3. et al., M.W.: JANUS 93: Towards Spontaneous Speech Translation. In: International Conference on Acoustics, Speech, and Signal Processing 1994. Adelaide, Australia (1994)
4. Amodei, D., Anubhai, R., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Chen, J., Chrzanowski, M., Coates, A., Diamos, G., et al.: Deep speech 2: End-to-end speech recognition in english and mandarin. arXiv preprint arXiv:1512.02595 (2015)
5. Caruana, R.: Multitask learning. Machine learning 28(1), 41–75 (1997)
6. Chen, D., Mak, B., Leung, C.C., Sivadas, S.: Joint acoustic modeling of triphones and trigraphemes by multi-task learning deep neural networks for low-resource speech recognition. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. pp. 5592–5596. IEEE (2014)
7. Collobert, R., Kavukcuoglu, K., Farabet, C.: Torch7: A Matlab-like Environment for Machine Learning. In: BigLearn, NIPS Workshop (2011)
8. Ghoshal, A., Swietojanski, P., Renals, S.: Multilingual training of Deep-Neural networks. In: Proceedings of the ICASSP. Vancouver, Canada (2013)
9. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd international conference on Machine learning. pp. 369–376. ACM (2006)
10. Gretter, R.: Euronews: A Multilingual Benchmark for ASR and LID. In: Fifteenth Annual Conference of the International Speech Communication Association (2014)
11. Grézl, F., Karafiát, M., Vesely, K.: Adaptation of multilingual stacked bottle-neck neural network structure for new language. In: Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on. pp. 7654–7658. IEEE (2014)
12. Heigold, G., Vanhoucke, V., Senior, A., Nguyen, P., Ranzato, M., Devin, M., Dean, J.: Multilingual Acoustic Models Using Distributed Deep Neural Networks. In: Proceedings of the ICASSP. Vancouver, Canada (May 2013)
13. Hinton, G.E., Osindero, S., Teh, Y.W.: A Fast Learning Algorithm for Deep Belief Nets. Neural computation 18(7), 1527–1554 (2006)
14. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735–1780 (1997)

15. Huang, H., Sim, K.C.: An Investigation of Augmenting Speaker Representations to Improve Speaker Normalisation for DNN-based Speech Recognition. In: ICASSP. pp. 4610–4613. IEEE (2015)
16. Kim, S., Hori, T., Watanabe, S.: Joint ctc-attention based end-to-end speech recognition using multi-task learning. arXiv preprint arXiv:1609.06773 (2016)
17. Laskowski, K., Heldner, M., Edlund, J.: The Fundamental Frequency Variation Spectrum. In: Proceedings of the 21st Swedish Phonetics Conference (Fonetik 2008). pp. 29–32. Gothenburg, Sweden (Jun 2008)
18. Lu, L., Kong, L., Dyer, C., Smith, N.A.: Multi-task learning with ctc and segmental crf for speech recognition. arXiv preprint arXiv:1702.06378 (2017)
19. Metze, F., Sheikh, Z., Waibel, A., Gehring, J., Kilgour, K., Nguyen, Q.B., Nguyen, V.H., et al.: Models of Tone for Tonal and Non-tonal Languages. In: Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on. pp. 261–266. IEEE (2013)
20. Miao, Y., Gowayyed, M., Metze, F.: 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. pp. 167–174. IEEE (2015)
21. Miao, Y., Zhang, H., Metze, F.: Towards Speaker Adaptive Training of Deep Neural Network Acoustic Models (2014)
22. Mohan, A., Rose, R.: Multi-lingual speech recognition with low-rank multi-task deep neural networks. In: Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on. pp. 4994–4998. IEEE (2015)
23. Müller, M., Stüker, S., Waibel, A.: Language Adaptive DNNs for Improved Low Resource Speech Recognition. In: Interspeech (2016)
24. Müller, M., Stüker, S., Waibel, A.: Language Feature Vectors for Resource Constraint Speech Recognition. In: Speech Communication; 12. ITG Symposium; Proceedings of. VDE (2016)
25. Müller, M., Waibel, A.: Using Language Adaptive Deep Neural Networks for Improved Multilingual Speech Recognition. IWSLT (2015)
26. Sak, H., Rao, K.: Multi-accent speech recognition with hierarchical grapheme based models (2017)
27. Saon, G., Kurata, G., Sercu, T., Audhkhasi, K., Thomas, S., Dimitriadis, D., Cui, X., Ramabhadran, B., Picheny, M., Lim, L.L., et al.: English Conversational Telephone Speech Recognition by Humans and Machines. arXiv preprint arXiv:1703.02136 (2017)
28. Saon, G., Soltau, H., Nahamoo, D., Picheny, M.: Speaker Adaptation of Neural Network Acoustic Models Using i-Vectors. In: ASRU. pp. 55–59. IEEE (2013)
29. Scanzio, S., Laface, P., Fissore, L., Gemello, R., Mana, F.: On the use of a multilingual neural network front-end. In: Proceedings of the Interspeech. pp. 2711–2714 (2008)
30. Schröder, M., Trouvain, J.: The German text-to-speech synthesis system MARY: A tool for research, development and teaching. International Journal of Speech Technology 6(4), 365–377 (2003)
31. Schubert, K.: Grundfrequenzverfolgung und deren Anwendung in der Spracherkennung. Master's thesis, Universität Karlsruhe (TH), Germany (1999), in German
32. Schultz, T., Waibel, A.: Fast bootstrapping of lvcsr systems with multilingual phoneme sets. In: Eurospeech (1997)
33. Schultz, T., Waibel, A.: Language-independent and language-adaptive acoustic modeling for speech recognition. Speech Communication 35(1), 31–51 (2001)

34. Soltau, H., Liao, H., Sak, H.: Neural speech recognizer: Acoustic-to-word lstm model for large vocabulary speech recognition. arXiv preprint arXiv:1610.09975 (2016)
35. Soltau, H., Metze, F., Fugen, C., Waibel, A.: A One-Pass Decoder Based on Polymorphic Linguistic Context Assignment. In: Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on. pp. 214–217. IEEE (2001)
36. Stüker, S.: Acoustic modelling for under-resourced languages. Ph.D. thesis, Karlsruhe, Univ., Diss., 2009 (2009)
37. Sutskever, I., Martens, J., Dahl, G., Hinton, G.: On the importance of initialization and momentum in deep learning. In: Proceedings of the 30th International Conference on Machine Learning (ICML-13). pp. 1139–1147 (2013)
38. Swietojanski, P., Ghoshal, A., Renals, S.: Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR. In: SLT. pp. 246–251. IEEE, IEEE (2012)
39. Vesely, K., Karafiat, M., Grezl, F., Janda, M., Egorova, E.: The language-independent bottleneck features. In: Proceedings of the Spoken Language Technology Workshop (SLT), 2012 IEEE. pp. 336–341. IEEE (2012)
40. Waibel, A., Hanazawa, T., Hinton, G., Shikano, K.: Phoneme Recognition Using Time-Delay Neural Networks. In: ATR Interpreting Telephony Research Laboratories (October 30 1987)
41. Wheatley, B., Kondo, K., Anderson, W., Muthusamy, Y.: An evaluation of cross-language adaptation for rapid hmm development in a new language. In: Acoustics, Speech, and Signal Processing, 1994. ICASSP-94., 1994 IEEE International Conference on. vol. 1, pp. I–237. IEEE (1994)
42. Xiong, W., Droppo, J., Huang, X., Seide, F., Seltzer, M., Stolcke, A., Yu, D., Zweig, G.: Achieving Human Parity in Conversational Speech Recognition. arXiv preprint arXiv:1610.05256 (2016)