Language Adaptive DNNs for Improved Low Resource Speech Recognition

Markus Müller, Sebastian Stüker, Alex Waibel

Institute for Anthropomatics and Robotics, Karlsruhe Institute of Technology, Karlsruhe, Germany

m.mueller@kit.edu, sebastian.stueker@kit.edu, alexander.waibel@kit.edu

Abstract

Deep Neural Network (DNN) acoustic models are commonly used in today's state-of-the-art speech recognition systems. As neural networks are a data driven method, the amount of available training data directly impacts the performance. In the past, several studies have shown that multilingual training of DNNs leads to improvements, especially in resource constrained tasks in which only limited training data in the target language is available.

Previous studies have shown speaker adaptation to be successfully performed on DNNs. This is achieved by adding speaker information (e.g. i-Vectors) as additional input features. Based on the idea of adding additional features, we here present a method for adding language information to the input features of the network. Preliminary experiments have shown improvements by providing supervised information about language identity to the network.

In this work, we extended this approach by training a neural network to encode language specific features. We extracted those features unsupervised and used them to provide additional cues to the DNN acoustic model during training. Our results show that augmenting acoustic input features with this language code enabled the network to better capture language specific peculiarities. This improved the performance of systems trained using data from multiple languages.

Index Terms: Multilingual acoustic modelling, neural networks, low-resource ASR

1. Introduction

Building Large Vocabulary Continuous Speech Recognition Systems (LVCSR) systems with decent performance requires a fair amount of training data. While sufficient data is available for languages like English, this is not the case for the majority of languages in the world. Therefore, it is challenging to build systems for those under resourced languages. The challenge is even bigger if no (transcribed) acoustic training data is available. Multiple techniques have been explored regarding the use of data from different languages to build a system for a particular target language. In the past, we explored different training strategies to train artificial neural networks (ANNs) in a multilingual fashion [1]. We showed, that using additional data from multiple languages is helpful for training neural networks that display a higher recognition accuracy.

Recently, we introduced language adaptive DNNs (LA-DNN)[2] which are able to better capture language specific peculiarities which resulted in a decrease of the word error rate (WER). By explicitly adapting the network to a certain language, the training data could be exploited in a more efficient way. In this work, we extended our approach by transitioning from an explicit to an implicit adaptation: Instead of modelling the language information (LID) directly, we trained a neural network to extract a language feature vector (LFV) that conveys language specific features. We also demonstrate, that these features carry a richer set of information than just the language identity. In addition to that, we show that this technique is also applicable to languages not seen during training.

For our experiments, we pretend English to be a low resource language. For building a LVCSR system we therefore restricted the amount of available data to 30h per language. In addition, we also evaluated our method for the task of phoneme boundary detection. In this scenario, we used a small dataset from Basaa, a sub-saharan African language. This data was only used for testing, we did not train or adapt our systems, hence this data and the language was not seen by our evaluated system.

This paper is organized as follows: In Section 2, we provide an overview of related work. In the next Section (3), we describe our proposed approach in detail. In Section 4 we describe our experimental setup, followed by the results in Section 5. We summarize our findings in the final Section 6 where we also provide an outlook to future work.

2. Related work

Nowadays, ANNs are a common part of LVCSR systems. Neural networks are being used in components like feature extraction, language modelling and acoustic modelling. In this work, we focus on the use of ANNs as part of the feature extraction pipeline as well as the acoustic modelling.

2.1. GMM Based Multilingual Systems

Before the widespread emergence of neural networks in LVCSR systems, using a GMM/HMM based approach for acoustic modelling was common. GMM/HMM systems are known to suffer in performance if trained multilingually. This problem of training multi- and crosslingual HMM/GMM systems has been addressed in the past. Techniques like ML-Mix or ML-Tag have been explored to exploit data from multiple languages for building GMM/HMM based systems [3]. There exist also techniques for building systems crosslingually [4].

2.2. Multilingual DBNFs

DNNs have shown to benefit from multitask learning [5]. The authors in [6] showed that the pre-training step is language independent. There exist multiple possibilities to use data from multiple languages during fine-tuning. One possibility is to share the hidden representations among different languages, but keep the output layers language specific ([7], [8], [9], [10]). Another possibility is the use of a global phoneme set [11]. For our approach, we used a global phoneme set to build a truly multilingual LVCSR system.

2.3. Augmenting Input Features

Many works demonstrated that the concept of augmenting the acoustic input features of neural networks with additional features increases the recognition performance of ASR systems. A common approach is to use i-Vectors [12] or Bottleneck Speaker Vectors (BSV) [13] to provide information about different speakers to the network. It is also possible to train a speaker adaptive neural network [14].

2.4. Phoneme Boundary Detection

Languages without writing systems or no available written language resources require different methods for building speech recognition systems. A first step towards such a system is the discovery of phoneme like units in order to build a system to perform phonetic transcriptions. For discovering such units, the audio has to be segmented. One approach is to detect acoustic changes in audio signals [15] in order to predict phoneme boundaries. To evaluate such boundaries, there exist different metrics [16]. The metric that we used to estimate the quality of the phoneme segmentation in this work is the F-Score, along with precision and recall.

3. Language Adaptive Deep Neural Networks

As outlined in the related work section, using resources from additional languages can increase the recognition performance of LVCSR systems if little or no data from the target language is available. We showed that augmenting the acoustic features with language identity (LID) information increases the system performance [2]. To add the LID to the acoustic features, we used a one hot encoding to create a feature vector with one dimension per language. But using this LID encoding has two drawbacks: The network has to be re-trained once a new language is added and the actual language has to be provided to the network in a supervised fashion. This approach requires the knowledge of the language used in order to give the correct LID information to the system.

In order to handle unseen languages, a new method is required. We propose a language feature vector (LFV) which is extracted using a DNN for building language adaptive DNNs (LA-DNNs). To extract this vector, we used an architecture similar to DBNFs. We used a feed-forward DNN with a bottleneck layer and took the output activations from that layer and used them as LFVs.

Figure 1 shows the network architecture. The setup consisted of two networks. The first network was used to extract DBNFs from acoustic input features. It was trained using a combination of lMel and tonal features with a context of 6 frames as input and CD phoneme states as targets. We trained it in a multilingual fashion featuring multiple output layers, one per language. The network featured 6 layers with 1000 neurons each. The second last layer was a bottleneck layer with a size of only 42 neurons.

The second stage network used the output of the BNF network as input. As we considered language properties to be long-term in nature, we used a larger context of 11 frames as input into the second network. To stretch the range of this context even further, we used only every n-th frame, thereby increasing the range n-fold by omitting frames in between. We determined the optimal context size in a series of experiments. The network was trained to determine the language identity using audio data from multiple languages. We added a bottleneck layer as second last hidden layer of this network. For the extraction of the language feature, we used the layers up until the bottleneck and discarded the other layers.

4. Experimental Setup

We evaluated our approach in a series of experiments. We used the Janus Recognition Toolkit (JRTk) [17] which features the IBIS single-pass decoder [18]. We trained our neural networks using a setup based on Theano [19]. For creation of the pronunciation dictionaries, we used MaryTTS [20].

4.1. Corpora

Our experiments were based on a speech corpus consisting of recordings from Euronews[1], a TV news station [21]. It consisted of approximately 70h of acoustic training data per language. The audio was sampled at 16 kHz. We used data from all 10 available languages—Arabic, English, French, German, Italian, Polish, Portuguese, Russian, Spanish and Turkish—as shown in Table 1. For testing, we used the provided English test set containing 37 recordings with a total length of 1.2h.

In addition to the Euronews corpus, we also used approx. 1h of audio from Basaa for the task of phoneme boundary detection. This dataset contained only one speaker and was recorded in a clean environment by re-speaking recordings that were originally recorded in the field. For details about this dataset, please refer to [22].

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

Table 1: Overview of the Euronews corpus

4.2. System Training

We carried out a first set of experiments using a combination of data from 6 languages—English, French, German, Italian, Russian and Turkish. The languages were selected based on both the availability of pronunciations from MaryTTS and data contained in the Euronews corpus. In this initial experiment, we selected 30h of data from each language on a per speaker basis. To bootstrap the initial models, we used a flat start approach. We built a GMM/HMM based context dependent (CD) system with 6,000 models.

4.3. DBNF Training

Based on this CD system, we extracted training data for training the DBNF network for the extraction of Bottleneck Fea-

[1] www.euronews.com

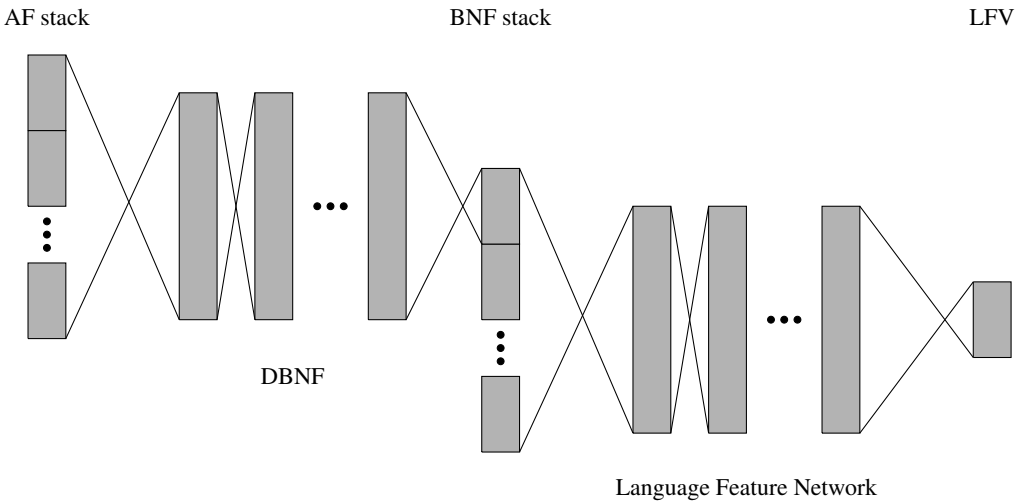

Figure 1: Overview of the network architecture used to extract language feature vectors (LFV). The acoustic features (AF) are being pre-processed in a DBNF in order to extract BNFs. These BNFs are being stacked and fed into the second network to extract LFVs.

tures (BNFs). As input features, we used a combination of lMEL, fundamental frequency variation (FFV) [23] and pitch [24] acoustic features. The use of tonal features had lead to improvements in combination with DNNs, even for non-tonal languages such as English [25]. We therefore included them as part of our default audio pre-processing pipeline. Those features were fed into the DBNF network using a context of 6 frames. In addition, we augmented the acoustic features with LFVs.

The network for the extraction of BNFs consisted of 6 hidden layers. The second last hidden layer was a bottleneck layer. While the other layers featured 1,000 neurons each, the bottleneck layer was very narrow, having only a size of 42 neurons. The network was layer-wise pre-trained using de-noising auto-encoders. For fine-tuning, we used stochastic gradient descent with newbob scheduling to adjust the learning rate. The output layers were trained on multilingual CD phoneme state targets.

4.4. Hybrid System Training

In order to obtain labels for training the hybrid system, we re-trained the GMM/HMM system using BNFs. Using this re-trained system, we obtained labels to train a second DNN with BNFs as input and CD phoneme states as targets. We stacked the BNF input features using a context of 7 and augmented them with LFVs. The DNN in our hybrid system featured 6 hidden layers with a size of 1600 neurons each. Similar to the DBNF training, we also performed layer wise pre-training followed by fine-tuning with newbob scheduling of the learning rate.

4.5. Language Feature Network Training

As outlined in section 3, network for extracting LFVs consisted of two separate networks. We therefore trained it in two steps. As we considered English to be a low resource language in this work, we did not use any English data during training. By not using any data from the target language, we demonstrate that LFVs are powerful enough to encode relevant features even for previously unseen languages.

In the first step, we trained the DBNF network using data from 5 languages—French, German, Italian, Russian and Turkish—70 hours each, resulting in 350 hours of acoustic training material in total. The BNFs extracted via this network were fed as input features into the second DNN. To train this second DNN, we used data from all 9 languages (except English) contained in the Euronews data set. It was trained to determine the language identity. As targets, we used the language identity encoded using one hot encoding. The output layer consisted of 9 neurons, one for each language.

We assumed the language information to be a more static feature in contrast to single phonemes. To capture this long term feature, we increased the context size of the network. For this, we carried out a preliminary set of experiments testing different context sizes to be fed into the language feature network. To evaluate the different context sizes, we divided the available data into a training and validation set with 10% of the data contained in the validation set and the rest in the training set. We measured the performance of the different context sizes using the classification error on the validation set.

4.6. Multilingual System

We evaluated our proposed method using a multilingual system built with data from 6 languages—English, French, German, Italian, Russian, Spanish and Turkish. We limited the amount of training data to 30h per language, resulting in 180h total. We trained a system without any language information as baseline. As contrastive experiments, we included numbers from systems trained using language identity information (LID). The evaluation was performed on English test data from Euronews. We used a 4-gram language model with a vocabulary of 100k words.

4.7. Crosslingual Phoneme Boundary Detection

We also evaluated our method by performing crosslingual phoneme boundary detection. We measured the accuracy of detected boundaries using unseen data from Basaa. In order to obtain phoneme boundaries as baseline, we used a multilingual system and adapted the acoustic models by doing a forced alignment using the phonetic transcripts of the Basaa recordings. We evaluated the F-score of the hypothesized phoneme boundaries

with respect to the baseline boundaries.

To determine the phoneme boundaries, we used a LVCSR system in a special configuration. It was trained in the same manner like the other systems, but we made some adjustments to detect phoneme boundaries. For this, we used a pronunciation dictionary containing only words consisting of single phonemes. The LFVs were fed into the DBNF. We used an unigram language model with equal probabilities for each phoneme. For evaluation, we used only the boundaries of the detected phonemes and discarded the phoneme identity information. A detailed description of this setup can be found in [22].

5. Results

In this section, we present the results that show improvements by the addition of LFVs to the acoustic input features. The system performance increases for both the LVCSR setup as well as the phoneme boundary detection.

5.1. Context width of LFVs

To determine the optimal context width, we performed a series of preliminary experiments. We carried them out using all available data from a subset of 6 languages—English, German, French, Italian, Russian and Turkish. As error measure we used the classification error on the validation set. Table 2 shows the classification performance for different context widths. The widths were varied using only very n-th frame, thereby increasing the spread and to cover a larger area while keeping the dimensionality of the input features identical. By using a context of 690ms, we observed a minimal validation error of 0.136. Using smaller or bigger context sizes did not lead to improvements.

Context width	Spread	Error
460ms	2	0.142
690ms	**3**	**0.136**
1380ms	6	0.139

Table 2: Overview of different context widths for LFV extraction

5.2. Multilingual System

We evaluated the use of language features as well as LID in a multilingual system setup. The results are shown in Table 3. The first row shows the WER of GMM/HMM systems with DBNF acoustic input features. Compared to the baseline, we observed a slight improvement by 3% relative in WER. The improvements for both LID and LFV are identical. As for the hybrid systems, by augmenting the acoustic input features with LID, we see an decrease of WER by 7.9% relative. The impact of LFVs is even bigger, resulting in a relative decrease of 9.3% of WER from 17.7% to 16.2% compared to using LID.

These results indicate that LFVs carry a richer set of information compared to LID alone. It is also important to note, that the LFVs were able to extract relevant language characteristics for English, even though the network was not trained on English data. This shows that LFVs generalize across languages and are not limited to the fixed set of languages they were trained on. By extracting LFVs on an unseen language, we also avoid the network taking advantage of non language related acoustic events that may be different across languages like specific jingles or music.

System	Baseline	LID	LFV
DBNF GMM/HMM	21.4%	20.7%	**20.7%**
DBNF Hybrid	19.1%	17.7%	**16.2%**
rel. improvement	—	7.9%	**9.3%**

Table 3: WER of multilingual systems, trained with LID or LFVs

5.3. Crosslingual Phoneme Boundary Detection

As last evaluation, we evaluated the accuracy of detected phoneme boundaries. We determined the accuracy by computing the F-Score and included numbers for precision and recall as well. Table 4 shows the results. In this scenario, we observed an increase in accuracy by adding LFVs to the acoustic input features of the network. Improvements can be observed for Precision, Recall and F-Score. This shows that LFVs increased the accuracy of the system in phoneme boundary prediction.

System	Baseline	with LFV
Precision	0.520	**0.542**
Recall	0.515	**0.532**
F-Score	0.518	**0.537**

Table 4: Overview of results for crosslingual phoneme segmentation. The F-Score shows 3.7% relative improvement.

6. Conclusion and Outlook

Building systems for languages with limited resources is a challenging task. The shortage in training data can be circumvented by using multi- and/or crosslingual modelling techniques. Similar to the adaptation of neural networks to different speakers, we have shown that it is possible to adapt networks to different languages. We proposed an approach towards improving the performance of ASR systems in low resource conditions. We investigated the use of LFVs as an additional source of information in combination with acoustic input features. Our experiments showed that LFVs are more powerful than the language identity information alone. Both evaluated tasks (LVCSR and phoneme boundary detection) benefitted from the addition of LFVs.

Our method enables DNNs to better learn and adapt to the characteristics of different languages. LFVs have also shown to be helpful on previously unseen languages. Future experiments include the optimization of hyper parameters of the DNN for LFV extraction. We would also like to investigate the use of LFVs to capture language specific properties even in monolingual environments. Similar to i-Vectors, LFVs could supply information about speaker specific language peculiarities to ANNs.

7. Acknowledgements

This work was realized in the framework of the ANR-DFG project BULB (ANR-14-CE35-002). The authors would like to thank the reviewers for their helpful comments.

8. References

[1] M. Müller, S. Stüker, Z. Sheik, F. Metze, and A. Waibel, "Multilingual deep bottle neck features - a study on language selection and training techniques," *Proceedings of the 11th International Workshop on Spoken Language Translation (IWSLT)*, 2014.

[2] M. Müller and A. Waibel, "Using language adaptive deep neural networks for improved multilingual speech recognition," *Proceedings of the 12th International Workshop on Spoken Language Translation (IWSLT)*, 2015.

[3] 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.

[4] S. Stüker, "Acoustic modelling for under-resourced languages," Ph.D. dissertation, Karlsruhe, Univ., Diss., 2009, 2009.

[5] R. Caruana, "Multitask learning," *Machine learning*, vol. 28, no. 1, pp. 41–75, 1997.

[6] P. Swietojanski, A. Ghoshal, and S. Renals, "Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR," in *Proceedings of the Spoken Language Technology Workshop (SLT), 2012 IEEE*, IEEE. IEEE, 2012, pp. 246–251.

[7] A. Ghoshal, P. Swietojanski, and S. Renals, "Multilingual training of Deep-Neural networks," in *Proceedings of the ICASSP*, Vancouver, Canada, 2013.

[8] S. Scanzio, P. Laface, L. Fissore, R. Gemello, and F. Mana, "On the use of a multilingual neural network front-end," in *Proceedings of the Interspeech*, 2008, pp. 2711–2714.

[9] G. Heigold, V. Vanhoucke, A. Senior, P. Nguyen, M. Ranzato, M. Devin, and J. Dean, "Multilingual Acoustic Models Using Distributed Deep Neural Networks," in *Proceedings of the ICASSP*, Vancouver, Canada, May 2013.

[10] K. Vesely, M. Karafiat, F. Grezl, M. Janda, and E. Egorova, "The language-independent bottleneck features," in *Proceedings of the Spoken Language Technology Workshop (SLT), 2012 IEEE*. IEEE, 2012, pp. 336–341.

[11] 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.

[12] G. Saon, H. Soltau, D. Nahamoo, and M. Picheny, "Speaker adaptation of neural network acoustic models using i-vectors," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 55–59.

[13] H. Huang and K. C. Sim, "An investigation of augmenting speaker representations to improve speaker normalisation for DNN-based speech recognition," in *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4610–4613.

[14] Y. Miao, H. Zhang, and F. Metze, "Towards speaker adaptive training of deep neural network acoustic models," 2014.

[15] O. Scharenborg, V. Wan, and M. Ernestus, "Unsupervised speech segmentation: An analysis of the hypothesized phone boundaries," *Acoustical Society of America, Journal of*, vol. 127, no. 2, pp. 1084–1095, 2009.

[16] Y. Qiao, N. Shimomura, and N. Minematsu, "Unsupervised optimal phoneme segmentation: objectives, algorithm and comparisons," in *Acoustics, Speech and Signal Processing (ICASSP), 2008 IEEE International Conference on*. IEEE, 2008, pp. 3989–3992.

[17] M. Woszczyna, N. Aoki-Waibel, F. D. Buø, N. Coccaro, K. Horiguchi, T. Kemp, A. Lavie, A. McNair, T. Polzin, I. Rogina, C. Rose, T. Schultz, B. Suhm, M. Tomita, and A. Waibel, "Janus 93: Towards spontaneous speech translation," in *International Conference on Acoustics, Speech, and Signal Processing 1994*, Adelaide, Australia, 1994.

[18] H. Soltau, F. Metze, C. Fugen, and A. Waibel, "A one-pass decoder based on polymorphic linguistic context assignment," in *Automatic Speech Recognition and Understanding, 2001. ASRU'01. IEEE Workshop on*. IEEE, 2001, pp. 214–217.

[19] Theano Development Team, "Theano: A Python framework for fast computation of mathematical expressions," *arXiv e-prints*, vol. abs/1605.02688, May 2016. [Online]. Available: http://arxiv.org/abs/1605.02688

[20] M. Schröder and J. Trouvain, "The german text-to-speech synthesis system mary: A tool for research, development and teaching," *International Journal of Speech Technology*, vol. 6, no. 4, pp. 365–377, 2003.

[21] R. Gretter, "Euronews: a multilingual benchmark for ASR and LID," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.

[22] M. Vetter, M. Müller, F. Hamlaoui, G. Neubig, S. Nakamura, S. Stüker, and A. Waibel, "Unsupervised phoneme segmentation of previously unseen languages," in *Proceedings of the Interspeech*, 2016.

[23] K. Laskowski, M. Heldner, and J. Edlund, "The Fundamental Frequency Variation Spectrum," in *Proceedings of the 21st Swedish Phonetics Conference (Fonetik 2008)*, Gothenburg, Sweden, Jun. 2008, pp. 29–32.

[24] K. Schubert, "Grundfrequenzverfolgung und deren Anwendung in der Spracherkennung," Master's thesis, Universität Karlsruhe (TH), Germany, 1999, in German.

[25] F. Metze, Z. Sheikh, A. Waibel, J. Gehring, K. Kilgour, Q. B. Nguyen, V. H. Nguyen *et al.*, "Models of tone for tonal and non-tonal languages," in *Automatic Speech Recognition and Understanding (ASRU), 2013 IEEE Workshop on*. IEEE, 2013, pp. 261–266.