

MULTILINGUAL ADAPTATION OF RNN BASED ASR SYSTEMS

Markus Müller, Sebastian Stüker, Alex Waibel

Institute for Anthropomatics and Robotics
Karlsruhe Institute of Technology, Karlsruhe, Germany
{m.mueller, sebastian.stueker, alexander.waibel}@kit.edu

ABSTRACT

In this work, we focus on multilingual systems based on recurrent neural networks (RNNs), trained using the Connectionist Temporal Classification (CTC) loss function. Using a multilingual set of acoustic units poses difficulties. To address this issue, we proposed Language Feature Vectors (LFVs) to train language adaptive multilingual systems. Language adaptation, in contrast to speaker adaptation, needs to be applied not only on the feature level, but also to deeper layers of the network. In this work, we therefore extended our previous approach by introducing a novel technique which we call “modulation”. Based on this method, we modulated the hidden layers of RNNs using LFVs. We evaluated this approach in both full and low resource conditions, as well as for grapheme and phone based systems. Lower error rates throughout the different conditions could be achieved by the use of the modulation.

Index Terms— Multilingual, automatic speech recognition, connectionist temporal classification, language feature vectors, low-resource

1. INTRODUCTION

Training multilingual speech recognition systems requires special methods. In low-resource conditions, training systems on data from multiple languages improves the performance. In a resource rich environment, using data from multiple languages often does not improve the performance, but might even affect it negatively. In both cases, adaptation techniques are required to improve the recognition accuracy and neural networks adapted to language characteristics have proven to perform better. This is similar to speaker adaptation, where adapted networks outperform unadapted ones.

However, language adaptation is more challenging than speaker adaptation: Collecting training data from several hundred speakers is possible. This amount of speakers enables networks to generalize upon speaker properties. For language adaptation, there are an order of magnitude less languages available than there are speakers. This renders generalization

across languages more difficult. Another factor is the task itself. When trained on data from multiple speakers of the same language, the same targets, e.g., phone states, are used. Different languages feature different, but sometimes overlapping, sets of targets. Although speech recognition for different languages are different tasks, they are related since all languages are being spoken by humans. This limits the sound inventory to the sounds that can be produced by the human vocal tract. Also, languages (from the same language family) potentially share sound inventories, as well as the set of targets the network is trained on.

Applying language adaptation techniques should therefore enable the networks to generalize better. Encoding language properties using, e.g., LFVs, like we showed in the past, allow networks to be trained language adaptive in such a way that they can exploit similarities and differences between languages. Unlike traditional GMM/HMM or DNN/HMM based systems, RNN/CTC based setups do not require explicit modelling of context dependent states which would then need to be adapted. Based on RNNs, these systems should be trained towards learning features based on language properties in order to be able to better perform in a multilingual scenario.

As we outlined in the related works section, several techniques for language adaptation have been proposed for traditional systems in the past. We proposed to use LFVs as additional input features for language adaptation. In this paper, we introduce a novel approach of integrating LFVs into recurrent network architectures based on the idea of Meta-PI networks. The effectiveness of our approach is demonstrated in a series of experiments, showing that the method presented here can be applied to both full- and low-resource conditions. In addition, we also omitted the pronunciation dictionary and built systems using graphemes only. In a multilingual scenario, this is particularly challenging as the network is required to learn pronunciations from multiple languages in parallel. To evaluate our systems, we use the token error rate (TER) as primary measure of the trained networks. But we also incorporated a RNN based language model (LM) for decoding to the determine the word error rate (WER).

This paper is organized as follows: In the next Section 2, we outlined related work in the field, followed by a de-

This work was realized in the framework of the ANR-DFG project BULB (ANR-14-CE35-002).

tailed description of the method proposed in Section 3. We described the experimental setup in Section 4, followed by the results of our experiments (Section 5). This paper concludes with Section 6, where we also outline possible future work.

2. RELATED WORK

2.1. Multi- and Crosslingual Speech Recognition Systems

Prior to the emergence of neural networks, ASR systems were typically built using a GMM/HMM based approach. Methods for training/adapting such systems cross- and multilingually were proposed to handle data sparsity [1, 2]. The process of clustering context-independent phones into context-dependent ones can also be adapted to account for cross- and multilinguality [3]. Due to their recurrent nature, RNNs are a powerful tool to model sequential dependencies, rendering the need for context-dependent targets superfluous. Using only context-independent targets has the advantage that no clustering is required.

2.2. Multilingual Bottleneck Features

In a resource constraint scenario, data from additional source languages are used to improve the performance. DNNs are typically trained in two steps: Pre-training and fine-tuning. It has been shown that the pre-training step is language independent [4]. The fine-tuning can be modified in multiple ways to account for additional languages. One approach includes the use of shared hidden layers, with language dependent output layers [5]. Combining multiple output layers into one is also possible [6].

2.3. Neural Network Adaptation

Feeding additional input features into a neural network is a common way for adaptation. A popular approach for speaker adaptation is to supply i-Vectors [7], which encode speaker characteristics in a low-dimensional representation. Speaker adaptive neural networks can be trained this way [8]. Such low dimensional codes can also be extracted using neural networks, called Bottleneck Speaker Vectors (BSVs) [9]. In the past, we proposed similar methods to adapt DNNs to multiple languages. We first introduced a method encoding the language identity using one-hot encoding [10]. We enhanced this method in a similar way to BSVs, by extracting Language Feature Vectors (LFVs) [11]. These vectors have shown to encode language properties instead of the language identity alone, even for languages not seen during training.

2.4. RNN Based ASR Systems

RNN based ASR systems are becoming increasingly popular. One method to train them is the use of the Connectionist tem-

poral classification (CTC) loss function [12], which does not require frame-level labels. It aligns a sequence of tokens automatically. As in traditional systems, phones, graphemes or both combined can be used as acoustic modeling units [13]. Given enough training data, even whole words can be used [14].

3. LANGUAGE ADAPTATION

In the past, we proposed methods for adapting multilingual neural network based ASR systems to languages using LFVs. Language Feature Vectors are a low-dimensional representation of language properties, extracted via a neural network. This network was trained to discriminate languages, based on log Mel and tonal features (FFV [15] and pitch [16]) typically used by ASR systems. A similar architecture as for extraction of BNFs was used. This architecture featured a bottleneck as second last layer. After training, the output activations of this layer were used as LFVs. To perform the adaptation, we appended LFVs to the acoustic features, similar to appending i-Vectors for speaker adaptation. In the results section, we included error rates using this method as contrastive experiments, denoted as “LFV app”. This method has shown to reduce error rates for multilingual GMM/HMM, DNN/HMM as well as RNN/CTC based systems.

3.1. Neural Modulation

Appending features for speaker adaptation to acoustic features is fitting, as changes in speaker characteristics are reflected within the signal. Multiple adaptation methods like VTLN or fMLLR which directly operate on the acoustic features were proposed. The same holds true for i-Vector based adaptation, where speaker adaptive systems can be trained to directly shift input features based on speaker properties [8]. But language properties are a higher order concept in contrast to speaker variations. Some aspects are based on acoustics, e.g. having the same phone in multiple languages, where a language specific coloring can be observed to some degree. But aspects like phonotactics or different sets of acoustic units require adaptation methods beyond the transformation of acoustic features. Here, adding features at deeper layers potentially enables better adaptation.

One possibility is a method first introduced as part of Meta-PI [17] networks. The key aspect is the use of Meta-PI connections, which allow to modulate the output of units by multiplication with a coefficient. Applied to language adaptation, we modulated the outputs of hidden layers with LFVs. Based on language features, the output of LSTM cells are attenuated or emphasized. This forces the cells in the hidden layer to learn or adapt to features based on language properties. Modulation can be considered related to dropout training [18], where connections are dropped on a random

basis. In the results section, we refer to this method as “LFV mod”.

We used a network configuration as shown in Figure 1. The basic architecture is inspired by Baidu’s Deepspeech 2. It combines two TDNN/CNN layers with 4 bi-directional LSTM layers. The output layer is a feed-forward layer which maps the output of the last LSTM layer to the targets. We chose the number of LSTM cells in each layer to be a multiple of the dimensionality of the LFVs. This way, we could structure the hidden layer into groups of LSTM cells containing an equal amount of units. The output of each group is then modulated with one dimension of the LFVs. The figure shows both configurations, “LFV app” and “LFV mod”, but only one method was applied at a time. In preliminary experiments, we determined modulating the output of the second LSTM layer to result in the best performance.

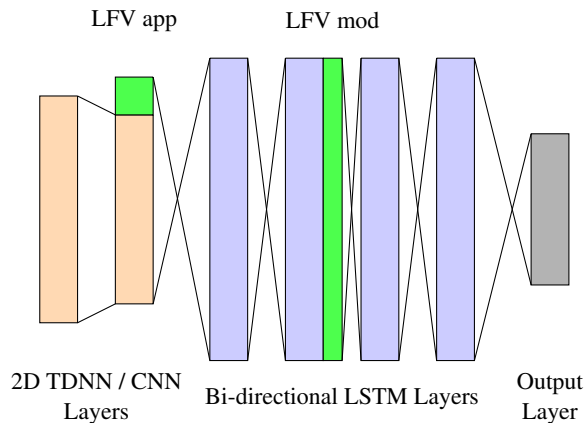


Fig. 1. Network architecture showing “LFV app”, as well as the proposed adaptation method “LFV mod”.

4. EXPERIMENTAL SETUP

We based our experiments on the Euronews corpus [19], which contains data from 10 languages. For each language, 70h of TV broadcast news recordings are available. For our experiments, we used a combination of 4 languages (English, French, German, Turkish), based on the availability of pronunciation dictionaries. We filtered utterances based on length, omitting very short ones ($< 1s$), and also removed ones having a transcript of more than 639 characters¹. Noises were only annotated in a very basic way with a single noise marker covering all different noise types, ranging from music, background and human noises. We therefore omitted utterances marked as noise. After applying all filtering, approx. 50h of data remained per language and was split into 45h of training and 5h of test data. For training, we created an

¹Internal limitation within the implementation of CUDA/warp-ctc, see: <https://github.com/baidu-research/warp-ctc>, accessed 2018-02-12

additional subset containing only 8h out of the 45h training set to evaluate our approach in a low-resource condition.

4.1. Acoustic Units

As acoustic units, we used both phones and graphemes. The pronunciation dictionaries were created using MaryTTS [20]. For merging the monolingual dictionaries, we mapped the phone-symbols to a multilingual phone set using the definition of articulatory features in MaryTTS’ language description files. In addition to systems based on phones, we also trained networks using graphemes as acoustic units. To indicate word boundaries, an additional token was used.

4.2. RNN/CTC Network Training

Multilingual Bottleneck Features (ML-BNFs) were used as input features. The ML-BNFs network was trained using data from 5 languages (French, German, Italian, Russian, Turkish). Input features to the network were log Mel and tonal features (FFV [15] and pitch [16]), extracted using a 32ms window with a 10ms frame-shift. The RNN network was trained using stochastic gradient descent (SGD) and Nesterov momentum [21] with a factor of 0.9. Mini-batch updates with a batch size of 15 were applied together with batch normalization. The utterances were sorted ascending by length to stabilize the training, as shorter utterances are easier to align.

4.3. Grapheme Based RNN LM

We used a RNN based LM, trained on graphemes as described in [22]. It featured 1 hidden layer with 1024 LSTM cells. The model was trained on only a very limited set of sentences, consisting of the training utterances of the acoustic model only. As language models are typically trained on several millions of sentences, this is not much training data. But the model should provide an indication whether the improvements observed as TER also result in a better word level speech recognition system.

4.4. Evaluation

We evaluated our proposed method varying two conditions: The availability of a pronunciation dictionary and the amount of data. An ASR system without language adaptation is used as baseline. First, we used the token error rate (TER) as primary measure to determine the performance without the use of external (language) models. For decoding, we use the same procedure as in [12] and greedily search for the best path. In addition to the TER, we also determined the word error rate (WER) using an RNN LM.

5. RESULTS

5.1. Grapheme Based Systems

First, we evaluated the use of graphemes as acoustic modeling units. We started using a network configuration with the RNN part having 420 LSTM cells per layer, trained using only 8h of data per language (see Table 1). Adding LFVs after the TDNN / CNN layers (“LFV app”) does lower the TER, but applying the method presented here (“LFV mod”) lowers the TER even more. Similar gains can be observed using the full

Condition	DE	EN	FR	TR
ML Baseline	30.8	38.0	29.4	30.9
LFV app	22.9	33.3	27.3	21.3
LFV mod	20.7	32.7	25.4	19.6

Table 1. TER of grapheme based system trained on 8h per language, 420 LSTM cells per layer

training set (Table 2). The use of more data lowered the TER, whereas the relative improvements were in the same order of magnitude. Training on more data also allowed for larger net-

Condition	DE	EN	FR	TR
Baseline	10.6	18.2	15.9	9.1
LFV app	9.5	16.1	14.3	8.1
LFV mod	9.1	15.5	13.6	8.0

Table 2. TER of grapheme based system trained on 45h per language, 420 LSTM cells per layer

works. In an additional experiment, we increased the number of LSTM cells per layer to 840. As shown in Table 3, the TER decreases in absolute terms, but the difference between addition and modulation becomes smaller.

Condition	DE	EN	FR	TR
Baseline	8.9	15.0	13.5	7.9
LFV app	7.9	13.6	11.8	7.1
LFV mod	7.7	13.3	11.7	7.1

Table 3. TER of grapheme based system trained on 45h per language, 840 LSTM cells per layer

5.2. Phoneme Based Systems

In the same notion as graphemes, we evaluated systems based on phonemes as acoustic modelling units. Starting with the limited data set (Table 4), improvements by the modulation (“LFV mod”) over the addition (“LFV app”) can be observed. Using all available training data and increasing the number

Condition	DE	EN	FR	TR
Baseline	21.7	27.2	23.9	21.6
LFV app	20.9	26.4	21.3	19.5
LFV mod	19.0	25.6	19.8	17.6

Table 4. TER of phoneme based system trained on 8h per language, 420 LSTM cells per layer

of LSTM cells per layer to 840, similar improvements could be achieved (Table 5). In contrast to the grapheme based setup (Table 3), modulating the layers (“LFV mod”) improves the performance over the simple addition (“LFV app”). The TERs of the grapheme based systems for German and Turkish are lower compared to their phone based counterparts. One reason for this is the quality of the pronunciation dictionary, which was created fully automatically based on a TTS system.

Condition	DE	EN	FR	TR
Baseline	9.6	14.6	12.1	8.5
LFV app	9.3	13.2	10.8	7.7
LFV mod	8.6	12.5	10.2	7.3

Table 5. TER of phoneme based system trained on 45h per language, 840 LSTM cells per layer

5.3. Decoding with RNN LM

To determine the WER, we ran a greedy decoding using a char based RNN LM on the English subset of the test data. The results shown in Table 6 indicate that the improvements of TER are also observable w. r. t. WER after decoding with a language model.

Setup	Baseline	LFV app	LFV mod
8h-420	32.4%	30.6%	29.9%
45h-840	29.2%	27.7%	27.3%

Table 6. WERs for English grapheme based systems.

6. CONCLUSION

Unlike speaker adaptation, where the collection of data covering hundreds of speakers is feasible, collecting data from that many languages is next to impossible. Optimizing the adaptation method is therefore key to maximize the performance in a multilingual scenario. We presented an improved method for language adaptation of RNNs in a multilingual setting. Modulating the outputs of a layer showed improvements over appending LFVs to input features.

7. REFERENCES

- [1] Tanja Schultz and Alex Waibel, “Fast bootstrapping of LVCSR systems with multilingual phoneme sets,” in *Eurospeech*, 1997.
- [2] Sebastian Stüker, *Acoustic modelling for under-resourced languages*, Ph.D. thesis, Karlsruhe, Univ., Diss., 2009, 2009.
- [3] Tanja Schultz and Alex Waibel, “Polyphone decision tree specialization for language adaptation,” in *Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2000, vol. 3, pp. 1707–1710.
- [4] Pawel Swietojanski, Arnab Ghoshal, and Steve Renals, “Unsupervised cross-lingual knowledge transfer in DNN-based LVCSR,” in *SLT*. IEEE, 2012, pp. 246–251, IEEE.
- [5] Karel Vesely, Martin Karafiat, Frantisek Grezl, Milos Janda, and Ekaterina Egorova, “The language-independent bottleneck features,” in *Proceedings of the Spoken Language Technology Workshop (SLT)*, 2012 IEEE. 2012, pp. 336–341, IEEE.
- [6] Frantisek Grézl, Martin Karafiát, and Karel Vesely, “Adaptation of multilingual stacked bottle-neck neural network structure for new language,” in *Acoustics, Speech, and Signal Processing (ICASSP)*. IEEE, 2014, pp. 7654–7658.
- [7] George Saon, Hagen Soltau, David Nahamoo, and Michael Picheny, “Speaker adaptation of neural network acoustic models using i-Vectors,” in *ASRU*. IEEE, 2013, pp. 55–59.
- [8] Yajie Miao, Hao Zhang, and Florian Metze, “Towards speaker adaptive training of deep neural network acoustic models,” 2014.
- [9] Hengguan Huang and Khe Chai Sim, “An investigation of augmenting speaker representations to improve speaker normalisation for dnn-based speech recognition,” in *ICASSP*. IEEE, 2015, pp. 4610–4613.
- [10] Markus Müller and Alex Waibel, “Using language adaptive deep neural networks for improved multilingual speech recognition,” *IWSLT*, 2015.
- [11] Markus Müller, Sebastian Stüker, and Alex Waibel, “Language adaptive DNNs for improved low resource speech recognition,” in *Interspeech*, 2016.
- [12] Alex Graves, Santiago Fernández, Faustino Gomez, and Jürgen Schmidhuber, “Connectionist temporal classification: labelling unsegmented sequence data with recurrent neural networks,” in *Proceedings of the 23rd international conference on Machine learning*. ACM, 2006, pp. 369–376.
- [13] Dongpeng Chen, Brian Mak, Cheung-Chi Leung, and Sunil Sivadas, “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)*. IEEE, 2014, pp. 5592–5596.
- [14] Hagen Soltau, Hank Liao, and Hasim Sak, “Neural speech recognizer: acoustic-to-word LSTM Model for large vocabulary speech recognition,” *arXiv preprint arXiv:1610.09975*, 2016.
- [15] Kornel Laskowski, Mattias Heldner, and Jens Edlund, “The fundamental frequency variation spectrum,” in *Proceedings of the 21st Swedish Phonetics Conference (Fonetik 2008)*, Gothenburg, Sweden, June 2008, pp. 29–32.
- [16] Kjell Schubert, “Grundfrequenzverfolgung und deren anwendung in der spracherkennung,” M.S. thesis, Universität Karlsruhe (TH), Germany, 1999, In German.
- [17] John B Hampshire and Alex Waibel, “The Meta-Pi network: Building distributed knowledge representations for robust multisource pattern recognition,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 7, pp. 751–769, 1992.
- [18] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov, “Improving neural networks by preventing co-adaptation of feature detectors,” *arXiv preprint arXiv:1207.0580*, 2012.
- [19] Roberto Gretter, “Euronews: A multilingual benchmark for ASR and LID,” in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [20] Marc Schröder and Jürgen 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] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton, “On the importance of initialization and momentum in deep learning,” in *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, 2013, pp. 1139–1147.
- [22] Thomas Zenkel, Ramon Sanabria, Florian Metze, Jan Niehues, Matthias Sperber, Sebastian Stüker, and Alex Waibel, “Comparison of decoding strategies for CTC acoustic models,” *arXiv preprint arXiv:1708.04469*, 2017.