MULTILINGUAL REGION-DEPENDENT TRANSFORMS

Martin Karafiát, Lukáš Burget, František Grézl, Karel Veselý and Jan "Honza" Černocký

Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Brno, Czech Republic
{karafiat,burget,grezl,iveselyk,cernocky}@fit.vutbr.cz

ABSTRACT

In recent years, trained feature extraction (FE) schemes based on neural networks have replaced or complemented traditional approaches in top performing systems. This paper deals with FE in multilingual scenarios with a target language with low amount of transcribed data. Continuing our previous work on multilingual training of Stacked Bottle-Neck Neural Network FE schemes, we concentrate on improving the discriminatively trained Region-Dependent Transforms. We show that multilingual training of RDT can be implemented by merging statistics from several languages. In our case we used up to 11 source languages to build a FE which generalize well for a new language. This allows us to build a strong bootstrapping model for the final ASR system. The results are produced on IARPA Babel data.

***Index Terms*—** Automatic speech recognition, Region-Dependent Transforms, Multilingual speech recognition, Feedforward neural networks

1. INTRODUCTION

Quick delivery of ASR system for a new language is one of the challenges in the community. Hand in hand with this requirement comes the limitation of available resources. Such scenario calls not only for automated construction of systems, that have been carefully designed and crafted "by hand" so far, but also for effective use of available resources. Unfortunately, the data collection and annotation is the most time- and money-consuming procedure. It naturally raises an idea to borrow the information from other sources. As all human beings share the same vocal tract architecture, automatic systems should be able to have the low-level components (feature extraction) built and trained on various sources of data.

ASR systems have been using a variety of transforms to adjust features, model parameters, or both, for better matching of the system to the target data. Among these, Region Dependent Transforms (RDT) [1, 2] are giving good performance due to discriminative training. Moreover they are effective in discriminative fusion of features. A typical front-end of our GMM system [9] consists of two stages:

1. Neural network (NN) based Bottle-Neck (BN) [3] features and standard spectral based features (PLP-HLDA) [9] are extracted from every speech frame.
2. The two streams of features are fused using discriminatively trained Region Dependent Transforms (RDT).

This paper investigates RDT training in the framework of IARPA BABEL, where the goal is to quickly train keyword spotting systems[1] for new languages with minimum in-domain resources. The program already encouraged the research in training multilingual systems and their porting to new languages [3, 4, 5]:

Our previous work [7] studied the possibility to train a multilingual NN, which would be able to extract features for a new language. Several approaches to create the target phoneme set for the multilingual training were explored. The best and safest approach was found in splitting the last softmax layer into several blocks where each block accommodates training targets from one language [8].

Note that the first study of portability of Neural Network (NN) based features was done in [6], where NNs trained on English data were applied to Mandarin and Levantine Arabic to produce probabilistic features. Consistent word error rate (WER) reduction was observed for both languages. Unlike in IARPA BABEL program, however, the amount of training data for each language was sufficient for training good neural networks (100 and 70 hours respectively).

This paper extends the idea of multilingual training also to training RDT. Some initial experiments were already presented in [9]. Fully language-independent multilingual feature extraction should generate a better starting point when the system is ported to a new language. Naturally, all of the trainable front-end blocks should be trained in multilingual fashion.

2. REGION DEPENDENT TRANSFORM

In the RDT framework, an ensemble of linear transformations is trained, typically using the discriminative Minimum Phone Error (MPE) criterion [10]. Each transformation corresponds to one region in partitioned feature space. Each feature vector is then transformed by a linear transformation corresponding to the region the vector belongs to. The resulting (generally nonlinear) transformation has the following form:

$$F_{RDT}(\mathbf{o}_t) = \sum_{r=1}^{N} \gamma_r(t)\mathbf{A}_r\mathbf{o}_t, \qquad (1)$$

where $\mathbf{A}_r$ is linear transformation corresponding to rth region, and $\gamma_r(t)$ is probability that feature vector $\mathbf{o}_t$ belongs to rth region.

This work was supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Defense US Army Research Laboratory contract number W911NF-12-C-0013. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoD/ARL, or the U.S. Government. The work was also supported by Technology Agency of the Czech Republic project No. TA04011311 "MINT" and IT4Innovations Centre of Excellence CZ.1.05/1.1.00/02.0070.

[1] Due to correlation between ASR and KWS system, this work evaluate a performance only on ASR level

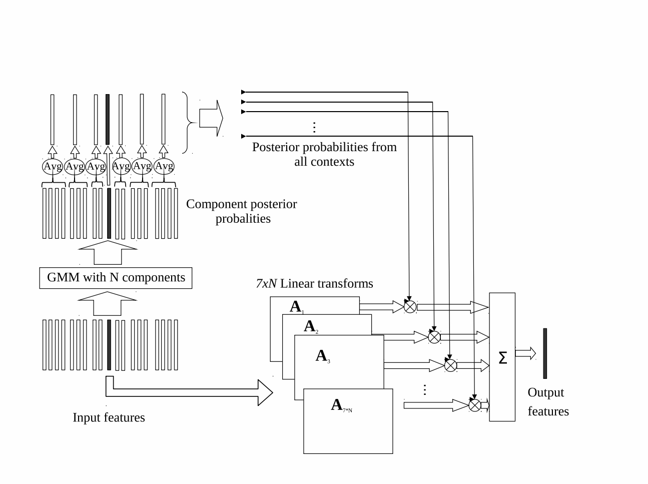

Fig. 1. *Region Dependent Transform.*

The probabilities $\gamma_r(t)$ are typically obtained using a GMM (pre-trained on the input features) as mixture component posterior probabilities. Usually, RDT parameters $\mathbf{A}_r$ and GMM-HMM acoustic model parameters are alternately updated in several iterations. While RDT parameters are updated using discriminative minimum phone-error (MPE) criterion, the acoustic model parameters are typically maximum likelihood (ML) trained on top of the features forwarded through the current RDT transformation [1],[2].

RDT can be seen as a generalization of proposed feature-MPE (fMPE) discriminative feature transformation. The special case of RDT with square matrices $\mathbf{A}_r$ was shown [2] to be equivalent to fMPE with offset features as described in [11]. From the fMPE recipe [1], we have adopted the idea of incorporating context information by considering $\gamma_r(t)$ corresponding not only to the current frame but also to the neighboring frames. From our experience, such incorporation of contextual information leads to significantly better results compared to the RDT style proposed in [2], where feature vectors of multiple frames were stacked at the RDT input. Therefore, our RDT configuration (figure 1) is very similar to the one described in the fMPE recipe [1].

2.1. Multilingual training of RDT

As was already mentioned, RDT parameters, $\mathbf{A}_r$, together with GMM-HMM acoustic model parameters are iteratively updated during RDT training. In our multilingual setting, however, we have several acoustic models one for each target language. The idea of multigual RDT training is to train single RDT model as a non-linear feature transformation, which is to be shared by all the acoustic models (or all the target speech recognizers). RDT model relies on the GMM defining the regions through $\gamma_r(t)$. In our experiment, we show that such GMM can be trained in multilingual fashion on data from several languages.

In our implementation, we use simple batch gradient descent algorithm for updating RDT parameters $\mathbf{A}_r$. Alternatively, L-BFGS algorithm can be used as suggested in [2]. From our experience, however, gradient descent usually provides better performing system (although its convergence is much slower). The gradients of the MPE objective with respect to RDT parameters $\mathbf{A}_r$ can be calculated as described in [2] (eq. 13). In our multilingual setting, gradients are calculated one for each target language on the training data and using the acoustic model of the corresponding language. All the language specific gradients are simply averaged into single gradient and used in a standard gradient descent algorithm to update the RDT parameters $\mathbf{A}_r$.

Y1 Langs. FLP hours	CA 65.0	PA 64.7	TU 56.6	TA 44.1	VI 53.2	
Y2 Langs FLP hours	AS 46.7	BE 53.6	HA 55.0	LA 71.6	ZU 57.8	Tam 72.7
Y3 Langs VLLP hours	TP 3.0					

Table 1. Amounts of data used for training.

Our experiments show that RDT trained in such multilingual fashion can not only be shared by all the target languages that were used for its training, but it can be also successfully used as a feature transformation for a new unseen language.

3. EXPERIMENTAL SETUP

3.1. Data

The IARPA Babel Program data simulates a case of what one could collect in limited time from a completely new language. It consists mainly of telephone conversation speech, but scripted recordings as well as far field recordings are present too.

The following language collection releases were used in this work (sorted by years of BABEL Program):

- Year 1 (Y1): Cantonese IARPA-babel101-v0.4c (CA), Pashto IARPA-babel104b-v0.4aY (PA), Turkish IARPA-babel105-v0.6 (TU), Tagalog IARPA-babel106-v0.2g (TA), Vietnamese IARPA-babel107b-v0.7 (VI)
- Year 2 (Y2): Assamese IARPA-babel102b-v0.5a (AS), Bengali IARPA-babel103b-v0.4b (BE), Haitian Creole IARPA-babel201b-v0.2b (HA), Lao IARPA-babel203b-v3.1a (LA), Zulu IARPA-babel206b-v0.1e (ZU), Tamil IARPA-babel204b-v1.1b (Tam)
- Year 3 (Y3): only TokPisin IARPA-babel207b-v1.0c (TP) was used.

Details about the languages can be found in [3]. Two main training scenarios were defined for each language – Full Language Pack (FLP), where all collected data was available for training – about 100 hours of speech; and Limited Language Pack (LLP) consisting only of one tenth of FLP. In year 3, the amount of training data for limited set was further decreased to size about 3h. This set is called Very Limited Language Pack (VLLP). In this condition, multilingual training and WEB text data collection were allowed on contrary to previous years where the acoustic model (AM) and language model (LM) training data were strictly bound to the respective Language Pack. Moreover, in year 3, pronunciation dictionaries were not provided and participants had to rely on graphemes in all conditions.

4. STRUCTURE OF ASR SYSTEMS

4.1. PLP system

Our speech recognition system was HMM based on cross-word tied-states triphones, it was trained from scratch using standard maximum likelihood training.

For the initial system, Mel-PLP features were generated (13 coefficients). Deltas, double- and triple-deltas were added, so that the

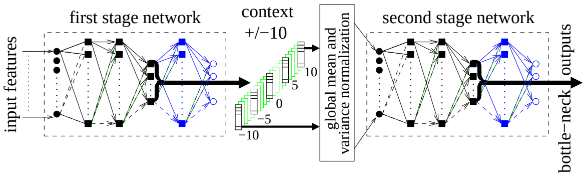

Fig. 2. Scheme of Stacked Bottle-Neck Neural Network feature extraction.

feature vector had 52 dimensions. Cepstral mean and variance normalization was applied with the means and variances estimated per conversation side. HLDA was estimated with Gaussian components as classes to reduce the dimensionality to 39. According to our previous experiments [9] the HLDA transform does not need to be trained on the target language, therefore the Tamil HLDA was selected for the further experiments.

4.2. Stacked Bottle-Neck feature extraction

The NN input features had 24 log Mel filter bank outputs concatenated with different fundamental frequency features: “BUT F0” had 2 coefficients (F0 and probability of voicing), “snack F0” was just a single F0 and “Kaldi F0” had 3 coefficients (Normalized F0 across sliding window, probability of voicing and delta). Fundamental frequency variation (FFV) had a 7 dimensional vector. Therefore, the whole feature vector had 24+2+1+3+7=37 coefficients (see [12] for details on fundamental frequency features).

Conversation-side based mean subtraction was applied and 11 consecutive frames were stacked. Hamming window followed by DCT consisting of 0th to 5th bases were applied on the time trajectory of each parameter resulting in 37×6=222 coefficients at the first-stage NN input (see Fig. 2).

The first-stage NN had four hidden layers with 1500 units each except the BN layer. The bottle-neck (BN) layer was the third hidden layer and its size is 80 neurons. Its outputs were stacked over 21 frames (+/-10) and down-sampled (every 5th is taken) before entering the second-stage NN. This NN had the same structure and sizes of hidden layers as the first-stage NN except for the BN layer with 30 neurons. The neurons in both BN layers had linear activation functions as they were reported to provide better performance [13].

The multilingual Stacked Bottle-Neck (SBN) NNs in this work were trained with the last layer – softmax – split into several blocks. Each block accommodates training targets from one language [8]. Context-independent phoneme states were used as the training targets. The NNs were trained on FLPs from Y1 + Y2 languages, excluding Tamil[2] (10 languages). The NN targets were monophone states obtained by forced alignment of training data with the initial PLP systems. Bottle-neck features generated in this way are further denoted “MultNN”.

5. EXPERIMENTAL RESULTS

5.1. Multilingual RDT on seen language

The RDT were trained on top of concatenated PLPHLDA+MultNN features stream. First, the initial language specific GMM-HMM acoustic models were estimated by Single-Pass-Retraining (SPR). The shared RDT transforms were initialized as identity matrices for $\mathbf{A}_r$ matrices corresponding to central frame and zero matrices otherwise (see Fig. 1). The RDT parameters and the new language specific acoustic models were iteratively trained until convergence as described in section 2.1.

Training data	GMM	WER[%]
Vietnam	Vietnam - 125G	53.3
Vietnam	Multilingual - 125G	53.3
Y1	Multilingual - 125	53.6
Y1	Multilingual - 500	53.7
Y1	Multilingual - 1000	53.9
Y1+Y2	Multilingual - 125G	**53.2**

Table 2. Multilingual RDT tested on Vietnamese FLP - one of the training languages.

Table 2 compares results obtained with multilingually trained RDT and RDT trined only on the Vietnamese target language. In this case, the system is trained on FLP, which means that there is already enough Vietnamese data to train RDT well. Therefore we do not expecting much improvement from the multilingual RDT training. Instead, we want to verify that the multilingual training generalizes well to any of the target languages without causing much performance degradation compared to the language specific training.

For the first two rows, the RDT parameter $\mathbf{A}_r$ are trained only on Vietnamese data, except that the GMM defining the regions (posteriors $\gamma_r(t)$) is trained on multilingual data in the second row. As the results are identical, we conclude that it is save to use the ”multilingual regions” in the following experiments.

The following lines shows results with all the RDT parameters trained in multilingual fashion. When training RDT only on Y1 data (5 languages) small degradation of 0.3% is observed when using the same number of regions (125 GMM components). As we are now effectively training on more data, we experimented with larger number of regions resulting in lager number of trainable RDT parameters (and more fine partitioning of the feature space). However, no gain was observed with the larger models compared to the configuration (125 components) that was also optimal for the monolingual training.

Training RDT on both Y1 and Y2 data gives 0.4% gain with respect to training only on Y1. Small improvement of 0.1% is obtained even over the language-specific system.

5.2. Application of multilingual RDT to unseen language

In the following experiments, we investigate how RDT trained on Y1 and Y2 languages generalizes to a new unseen language, where amount of training data is severally limited. System for TokPisin (one of the languages from Y3 languages) is trained on Very Limited Language Pack (VLLP) which was about 3h of data. The RDT input features (PLPHLDA+MultNN) were processed through RDT pretrained on Y1+Y2 data The initial PLP based GMM-HMM system was retrained using SPR on the RDT transformed features.

The first two rows of Table 3 shows that training RDT on the target language data give better performance compared to borrowing the pre-trained RDT from a single different language even if the amount of the training data is much smaller (more than an order of magnitude less data). On the other hand, use of the multilingual RDTs results to significant gains. Relatively small gain (0.1% absolute) was obtained form training RDT on 11 (Y1+Y2) languages compared to using only 5 (Y1) languages. The configuration with 125 GMM regions and RDT trained on 11 languages is used in the

[2] it was the last language delivered in Y2, therefore significant portion of the work was done without it.

Training data	GMM	WER[%]
TokPisin VLLP	TokPisin VLLP - 125G	51.2
Tagalog FLP	Tagalog FLP - 125G	52.0
Y1	Multilingual - 125G	50.3
Y1+Y2	Multilingual - 125G	**50.2**

Table 3. Multilingual RDT on TokPisin.

Initial features	Unadapted ML	Full system
PLP	70.0	44.9
MultNN - 10Lang	51.9	44.1
MultRDT	**50.0**	**43.7**

Table 4. WER [%] obtained with different flat-start features on TokPisin language.

following experiments and denoted as "MultRDT".

5.3. Building the complete system

Table 4 shows results for our experiment, where different features were used to train the TokPisin VLLP HMM system from scratch. The system only differ in the features used for their traing. Otherwise the whole system architecture and training procedure is the same. The first column in the table corresponds to ML trained unadapted system, which is obtained in an early stage of training the full system. The second column shows results for the full speaker adapted MPE trained GMM-HMM ststem making also use of the additional more advanced techniques:

- Fine-Tuning of Multilingual NN on target language data [3]
- Semi-Supervised Training of NN to deal with the problem of small amount of transcribed data (about 70h of the data can be used for unsupervised training) [16].
- The second stage NN of the SBN architecture (see Fig. 2) was trained on speaker-adapted features [12].

In the case of the unadapted ML trained system, we obtain very poor performance with the simple PLP features. Large improvement (18% absolute) is obtained when the previously proposed MultNN features are used. Additional 1.9% absolute improvement is obtain from applying the pre-trained multilingual RDT transformation on top of PLPHLDA+MultNN features. The NN based features and RDT transforms are trained to learn acoustics clues, therefore straightforward Maximum Likelihood (ML) flat-start models performs significantly better than traditional spectrum-based PLP features. Moreover, the NN-based features provides significantly faster convergence with less Gaussians due to the emergence of articulatory clusters [14].

As expected, smaller gains are observed for the full adapted and discriminatively trained system. Still, the best performance is obtained with the features extracted using the pretrained multilingual BN features and multilingual RDT transform. Note also that it is very easy and fast to train the unadapted ML system (training takes few hours) compared to the full system, which takes days to train.

6. CONCLUSION

This paper presented our further steps in the development of a feature extraction scheme easily transferable to a new language with severely limited training data. In addition to multilingual training of the bottle-neck neural networks explored previously, we have shown that similarly trained RDT can be beneficial for adapting the system to the target language and domain and for discriminative fusion of complementary feature streams. It is encouraging to see that the advantages brought by RDT do not vanish in the full system including a variety of techniques pushing the performance up, but that multlingual RDT initialization still contributes a solid 0.4% absolute improvement.

7. REFERENCES

[1] Daniel Povey, Brian Kingsbury, Lidia Mangu, George Saon, Hagen Soltau, and Geoffrey Zweig, "fMPE: Discriminatively trained features for speech recognition," in *in Proc. IEEE ICASSP*, 2005.

[2] Bing Zhang, Spyros Matsoukas, and Richard Schwartz, "Recent progress on the discriminative region-dependent transform for speech feature extraction," in *Proc. of Interspeech 2006*, Pittsburgh, PA, USA, Sep 2006, pp. 2977–2980.

[3] František Grézl and Martin Karafiát, "Adapting multilingual neural network hierarchy to a new language," in *Proceedings of the 4th International Workshop on Spoken Language Technologies for Under- resourced Languages SLTU-2014. St. Petersburg, Russia, 2014*. 2014, pp. 39–45, International Speech Communication Association.

[4] Zoltan Tuske, David Nolden, Ralf Schluter, and Hermann Ney, "Multilingual MRASTA features for low-resource keyword search and speech recognition systems," in *Proc. of Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, Florence, Italy, May 2014, IEEE, pp. 5607–5611.

[5] Quoc Bao Nguyen, Jonas Gehring, Markus Muller, Sebastian Stuker, and Alex Waibel, "Multilingual shifting deep bottleneck features for low-resource ASR," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, Florence, Italy, 2014, IEEE, pp. 5607–5611.

[6] A. Stolcke, F. Grézl, M.Y. Hwang, X. Lei, N. Morgan, and D. Vergyri, "Cross-domain and cross-language portability of acoustic features estimated by multilayer perceptrons," in *Proceedings of ICASSP 2006*, Toulouse, FR, 2006, pp. 321–324.

[7] F. Grézl, M. Karafiát, and M Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proceedings of ASRU 2011*, 2011, pp. 359–364.

[8] Karel Veselý, Martin Karafiát, František Grézl, Miloš Janda, and Ekaterina Egorova, "The language-independent bottleneck features," in *Proceedings of IEEE 2012 Workshop on Spoken Language Technology*. 2012, pp. 336–341, IEEE Signal Processing Society.

[9] Martin Karafiát, Miloš Janda, Jan Černocký, and Lukáš Burget, "Region dependent linear transforms in multilingual speech recognition," in *Proc. International Conference on Acoustics, Speech, and Signal Processing 2012*. 2012, pp. 4885–4888, IEEE Signal Processing Society.

[10] D. Povey, *Discriminative training for large vocabulary speech recognition*, Ph.D. thesis, University of Cambridge, 2003.

[11] Daniel Povey, "Improvements to fMPE for discriminative training of features," in *Proc. of Interspeech2005*, Lisbon, Portugal, Sep 2005, pp. 2977–2980.

[12] Martin Karafiát, František Grézl, Mirko Hannemann, Karel Veselý, Igor Szoke, and Jan ”Honza” Černocký, “BUT 2014 Babel system: Analysis of adaptation in NN based systems,” in *Proceedings of Interspeech 2014*, Singapure, September 2014, IEEE.

[13] Karel Veselý, Martin Karafiát, and František Grézl, “Convolutive bottleneck network features for LVCSR,” in *Proceedings of ASRU 2011*, 2011, pp. 42–47.

[14] Ngoc Thang Vu, Jochen Weiner, and Tanja Schultz, “Investigating the learning effect of multilingual bottle-neck features for ASR,” in *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, 2014, pp. 825–829.

[15] Martin Karafiát, Karel Veselý, Igor Szőke, Lukáš Burget, František Grézl, Mirko Hannemann, and Jan Černocký, “BUT ASR system for BABEL surprise evaluation 2014,” in *Proceedings of 2014 Spoken Language Technology Workshop*. 2014, pp. 501–506, IEEE Signal Processing Society.

[16] Grezl F., Karafiat M., and Vesely K., “Adaptation of neural network feature extractor for new language,” in *in Proceedings of ASRU 2013*, Olomouc, Czech Republic, 2013.