REGION DEPENDENT LINEAR TRANSFORMS IN MULTILINGUAL SPEECH RECOGNITION

Martin Karafiát[1], *Miloš Janda*[1], *Jan Černocký*[1] *and Lukáš Burget*[2,1]

(1) Brno University of Technology, Speech@FIT, Czech Republic
(2) SRI International, Menlo Park, CA, USA

ABSTRACT

In today's speech recognition systems, linear or nonlinear transformations are usually applied to post-process speech features forming input to HMM based acoustic models. In this work, we experiment with three popular transforms: HLDA, MPE-HLDA and Region Dependent Linear Transforms (RDLT), which are trained jointly with the acoustic model to extract maximum of the discriminative information from the raw features and to represent it in a form suitable for the following GMM-HMM based acoustic model. We focus on multi-lingual environments, where limited resources are available for training recognizers of many languages. Using data from GlobalPhone database, we show that, under such restrictive conditions, the feature transformations can be advantageously shared across languages and robustly trained using data from several languages.

Index Terms— HLDA, Region Dependent Transforms, Minimum Phone Error, fMPE, multilingual speech recognition

1. INTRODUCTION

Building speech-to-text systems on limited amount of data is shifting to the center of interest of speech recognition community. The main problem is the need of data for acoustic model training. Several techniques have been investigated to cope with this problem, such as cross-language transfer, language adaptation, bootstrapping, multilingual systems, or recently introduced Subspace Gaussian Mixture models [1]. All these techniques build acoustics models on top of a fixed feature extraction scheme. However, state-of-the-art speech recognition systems usually use advanced feature extraction techniques, where transformations are trained on data to post-process raw features. Training the transformations on a specific data can make the final features language- or channel-dependent. Typical examples of such transformations are linear transformations such as HLDA [2] or MPE-HLDA [3] and non-linear ones such as Region Dependent Transforms [4] or neural networks (NN) [5, 6].

The first study of portability of NN-based features was done in [5] and our previous work [6] focused on behavior of probabilistic and bottle-neck features trained on a particular language in a system designed for a different language. The conclusion was that NN-features do not generalize well when applied for different languages. Our approaches to obtaining language-independent NN features described in [6] was to train a single NN on data from multiple languages. Similarly to the transformations that will be described later in this paper, the resulting NN functioned as language-independent transformation of raw features. Training the NN in this fashion is, however, not very elegant — it is trained for frame-by-frame classification of phones of multiple languages, and ad-hoc decision has to be made on how such classes are defined (e.g. multilingual phone set based on IPA table). Also, the NNs are trained independently of the final language-dependent speech recognition systems.

On the other hand, Heteroscedastic Linear Discriminant Analysis (HLDA) or discriminatively trained Region Dependent Linear Transforms (RDLT) (that are the feature transformations that we experiments with in this work) are optimized for a specific acoustic model. In this paper, we show how these transformations can be estimated on data from several languages using multiple language-specific acoustic models. Such transformation can be shared across the language-specific speech recognition systems, and lead to language-independent feature extraction scheme, that is highly desired while developing (or rapidly prototyping) recognition systems for new languages. Another advantage is, that sharing transformations instead of models (as it is now common for multilingual systems), does not require merging of phoneme inventories.

2. HETEROSCEDASTIC LINEAR DISCRIMINANT ANALYSIS

HLDA [2] is a technique for estimating linear transformation:

$$F_{HLDA}(\mathbf{o}_t) = \mathbf{A}\mathbf{o}_t, \tag{1}$$

where $\mathbf{o}_t$ is input feature vector at time t, and $\mathbf{A}$ is transformation matrix. HLDA transformation allows us to reduce feature dimensionality while preserving information important for discrimination between classes. It also rotates the features to make them suitable for models, where the feature distribution is modeled using mixtures of Gaussians with diagonal covariance matrices. We use efficient iterative algorithm [7] to estimate matrix $\mathbf{A}$. It requires mean, count of occurrences and full-covariance matrix for each class. In our experiments, the classes are defined by each Gaussian mixture component. The selection, that feature vector $\mathbf{o}_t$ belongs to class j, is given by the value of occupation probability $\gamma_j(t)$.

HLDA in multilingual mode: The means, counts and covariance matrices are the only needed statistics to estimate the transformation. Therefore, the statistics are collected by language-specific HMM systems and stacked. Next, the shared HLDA transformation is estimated and cloned to all language-specific systems. New HMM models are re-estimated by single-pass retraining.

3. REGION DEPENDENT LINEAR TRANSFORMS

In the RDLT framework, an ensemble of linear transformations is trained typically using Minimum Phone Error (MPE) criterion [8].

This work was partly supported by Czech Ministry of Trade and Commerce project No. FR-TI1/034, Technology Agency of the Czech Republic grant No. TA01011328, Czech Ministry of Education project No. MSM0021630528. Martin Karafiát was supported by Grant Agency of the Czech Republic post-doctoral project No. P202/12/P400.

Each transformation corresponds to one region in partitioned feature space. Each feature vector is then transformed by a linear transformation corresponding to the region that the vector belongs to. The resulting (generally nonlinear) transformation has the following form:

$$F_{RDLT}(\mathbf{o}_t) = \sum_{r=1}^{N} \gamma_r(t)(\mathbf{A}_r \mathbf{o}_t + \mathbf{b}_r), \tag{2}$$

where $\mathbf{A}_r$ and $\mathbf{b}_r$ are linear transformation and bias corresponding to rth region and $\gamma_r(t)$ is probability that feature vector $\mathbf{o}_t$ belongs to rth region. The probabilities $\gamma_r(t)$ are typically obtained using Gaussian Mixture Model GMM (pre-trained on the input features) as mixture component posterior probabilities. Usually, RDLT parameters $\mathbf{A}_r$, $\mathbf{b}_r$ and ASR acoustic model parameters are alternately updated in several iterations. While RDLT parameters are updated using discriminative MPE criterion, ML update is typically used for acoustic model parameters. As proposed in [9] and described in RDLT context in [4], ML update of acoustic model parameters has to be taken into account when optimizing RDLT parameters (see indirect derivatives in [9]). Otherwise, the discriminative power obtained from MPE training of RDLT feature transformation is mostly lost after ML acoustic model re-training.

In our experiments, we closely followed the training recipe described in [4]. We do not use the bias terms $\mathbf{b}_r$ (the number of their parameters would anyway be only a small proportion of parameters in matrices $\mathbf{A}_r$). In agreement with results reported in [4], we have found that omitting the bias terms has little effect on the performance.

RDLT can be seen as a generalization of previously proposed fMPE discriminative feature transformation. The special case of RDLT with square matrices $\mathbf{A}_r$ (i.e. without dimensionality reduction of input features) was shown [4] to be equivalent to fMPE with offset features as described in [10]. This is also the configuration used in our experiments. From the fMPE recipe [9], we have also 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 (see section 3 for more details). From our experience, such incorporation of contextual information leads to significantly better results compared to the RDLT style proposed in [4], where feature vectors of multiple frames were stacked at the RDLT input and transformations with dimensionality reduction were used to recover the original feature dimensionality. Therefore, our RDLT baseline system configuration is very similar to the one described in the fMPE recipe.

In the trivial case, where all feature frames are considered to belong to only one single region, RDLT comprises only one discriminatively trained linear transformation. This configuration is also know as Discriminative HLDA [3] or MPE-HLDA:

$$F_{MPE-HLDA}(\mathbf{o}_t) = \mathbf{A}\mathbf{o}_t \tag{3}$$

RDLT in multilingual mode: RDLT are estimated using gradient-descent algorithm. Statistics needed to compute derivatives of MPE objective function are well described in [4] (equations 19–21). To obtain shared update statistics, it is enough to sum statistics from language-specific speech recognition systems. Computing derivatives and estimation of the new RDLT update follows the standard procedure. It is an iterative procedure, so the updated transformation is cloned to all language-specific systems, new HMM models are re-estimated and new statistics are collected. This process is repeated till convergence is reached.

Lang.	Speakers	TRAIN	TEST
GE	77	17	1.3
CZ	102	27	1.9
EN	311	15	1.0
SP	100	21	1.2
PO	102	27	1.0
TU	100	15	1.4
VN	129	16	1.3
RU	115	20	1.4
CTS	5446	270	-

Table 1. Numbers of speakers and amounts of audio material.

4. EXPERIMENTAL SETUP

4.1. Data

The data comes from multilingual database GlobalPhone [11]. The database covers 19 languages with an average of 20 hours of speech from about 100 native speakers per language. This database aims for an acceptable Out Of Vocabulary (OOV) rate in test sets, and contains newspaper articles read by native speakers (both genders, ages 18–81 years). Speech was recorded in office-like environment by high quality equipment. We converted the recordings to 8kHz, 16 bit, mono format.

The following languages were selected for the experiments: Czech (CZ), German (GE), Portuguese (PO), Russian (RU), Spanish (SP), Turkish (TU) and Vietnamese (VN). These languages were complemented with English (EN) taken from Wall Street Journal database. See Tab. 1 for detailed numbers of speakers and data partitioning. Each individual speaker appears only in one set. The partitioning followed the GlobalPhone recommendation.

When preparing the databases, several problems were encountered. The biggest issue was the low quality of dictionaries with many missing words. The Vietnamese dictionary was missing completely. The typos and miss-spelled words were corrected, abbreviations were expanded and missing pronunciations were generated with in-house grapheme-to-phoneme conversion tool. The dictionaries for Vietnamese and Russian were obtained from Lingea[1]. The CMU dictionary was used for English. Each language has its own phoneme set.

The data for Language Model (LM) were obtained from Internet sources (newspaper articles) using `RLAT` and `SPICE` tools from the KIT/CMU team[2]. The sizes of corpora gathered for LM training, and the sources are given in Tab. 2. Bigram LMs were generated for all languages except Vietnamese — a syllable language — for which a trigram LM was created.

4.2. Large data

Since the data in our database is quite small (about 15 hours for training) we were interested what would be the performance with a larger training set. For this purpose, Switchboard and CallHome English corpora were used. This data is Conversational Telephone Speech (CTS) which presents a different speaking style from read speech in GlobalPhone. Additional differences are due to the technical parameters of the recordings: telephone channel causes band limitation and adds noise to speech signal. These differences are another subject of interest – will the large amount of training data have positive effect

[1] http://www.lingea.com

[2] http://i19pc5.ira.uka.de/rlat-dev, http://plan.is.cs.cmu.edu/Spice

Lang	OOV	Dict Size	LM Corpus Size	WWW Server
GE	1.92	375k	19M	www.faz.net
CZ	3.08	323k	7M	www.novinky.cz
EN	2.30	20k	39M	WSJ - LDC2000T43
SP	3.10	135k	18M	www.aldia.cr
PO	0.92	205k	23M	www.linguateca.pt/cetenfolha
TU	2.60	579k	15M	www.zaman.com.tr
VN	0.02	16k	6M	www.tintuconline.vn
RU	1.44	485k	19M	www.pravda.ru

Table 2. OOV rates, dictionary sizes, LM sizes and sources for individual languages.

-	no HLDA	lang-dep.	EN	8L	8L+CTS
EN	17.6	**16.8**	**16.8**	17.1	**16.8**
GE	28.2	27.2	27.0	**26.9**	27.0
SP	25.0	23.6	23.6	**23.5**	23.8
PO	28.0	26.8	**26.3**	26.5	26.7
TU	34.6	32.3	**32.1**	32.4	33.0
VI	28.5	**25.0**	25.6	25.3	25.7
RU	35.4	33.0	**32.9**	**32.9**	33.4
CZ	24.3	23.0	**22.5**	22.6	22.6

Table 3. WER of HLDA systems: no HLDA system, lang-dep. - language-specific HLDA training, EN - HLDA estimated on English, 8L - HLDA estimated on 8 languages, 8L+CTS - HLDA estimated on 8 languages plus CTS data.

on system performance or will it be outweighed by the difference between training and test data?

4.3. Recognition systems

Speech recognition systems are HMM-based cross-word tied-states triphones, with approximately 3000 tied states and 18 Gaussian mixtures per state, trained from scratch using mix-up maximum likelihood training. The features are 13 Mel-Frequency PLP coefficients augmented with their deltas, double-deltas and for HLDA, also triple-deltas. Cepstral mean and variance normalization is applied with the mean and variance vectors estimated on each conversation side. HLDA is estimated with Gaussian components as classes and the dimensionality is reduced from 52 to 39.

5. HLDA EXPERIMENTS

Table 3 shows Word Error Rates (WER) of HLDA systems. The first and second columns present baseline systems without HLDA and with transformation estimated for each particular language. “HLDA EN” shows system performance with HLDA estimated by English system and used in other ones. It is interesting to see that HLDA is not too language-dependent. For most of the languages, English HLDA even outperforms the transformation estimated by corresponding system. The last two columns present the results with HLDA estimated from merged statistics. The results are close to baseline so, merging statistics does not have any significant effect. Adding telephone data also did not bring improvement for any language (including English). This was probably caused by different channels in GlobalPhone and CTS.

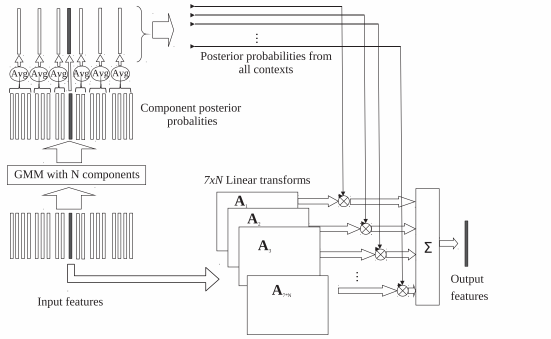

Fig. 1. RDLT with context transformations.

6. RDLT EXPERIMENTS

Different configurations of RDLT applied in multilingual scenario were examined: we began with already mentioned trivial case, where all feature frames are considered to belong to only one single region (MPE-HLDA). Next, we experimented with more sophisticated versions where information about regions from the current and context frames is incorporated: posterior probabilities of the GMM components for the current frame are stacked with the averages of posteriors for adjacent frames 1-2, 3-5 and 6-9 on the right and likewise on the left (i.e. 7 groups spanning 19 context frames in total). The resulting 7000-dimensional vector served as weights $\gamma_r(t)$ in (2) corresponding to 7000 transformations (39×39 matrices). Block diagram demonstrating such RDLT configuration is shown in Figure 1. In [12], we presented significant gain by adding such posterior probabilities from adjacent frames.

The GMM model is created by pooling and merging all Gaussian components from well trained baseline ML models. More details about the clustering algorithm can be found in [1].

6.1. MPE-HLDA

Since MPE-HLDA transformations are discriminatively trained to minimize phone error, they should be more language-dependent than ML-trained ones as each language contains different phonetic classes.

Table 4 show results of systems with one single transformation (MPE-HLDA). Standard HLDA estimated from all languages (“HLDA 8L” in table 3) was used as a starting point in all experiments to ensure that all results are comparable. The first and second columns show a 0.0%-0.4% drop in accuracy when a transformation trained on a different language (English) was used. When the transformation is shared across all languages, the drop is between 0%-0.1% (except for Turkish and Russian). Therefore, sharing produces more or less robust estimates. However, language-dependent transformations still provide the best performance. Adding telephone speech helps only to English set due to adding data from same language.

6.2. RDLT experiments

The improvements from discriminative training are quite small in the case of single MPE-HLDA transformation as there are not many parameters to train. Significant improvements can be, however, obtained with full RDLT, where an ensemble of linear transformations

-	lang-dep.	EN only	8L	8L+CTS
EN	16.6	16.6	16.7	**16.3**
GE	**26.6**	26.9	26.8	27.1
SP	**23.1**	23.3	23.2	23.3
PO	**26.1**	**26.1**	**26.1**	26.4
TU	**32.0**	32.2	32.6	32.6
VI	**24.9**	25.3	25.3	25.1
RU	**32.3**	**32.3**	32.6	32.4
CZ	**22.3**	22.5	22.5	22.6

Table 4. Accuracies of MPE-HLDA systems. lang-dep - language-specific transformation training, EN only - transformation trained on English, 8L - transformation trained on all 8 languages, 8L+CTS - transformation trained on all 8 languages plus CTS.

System	lang-dep	EN only	8L w/o VI	8L	8L+CTS
EN	15.1	15.1	14.7	14.6	**14.2**
GE	24.7	26.2	**24.2**	24.4	24.6
SP	21.8	23.5	21.9	**21.6**	22.1
PO	23.8	25.3	**23.3**	23.4	23.7
TU	29.7	31.4	**29.6**	29.8	30.0
VI	**22.6**	25.8	24.4	23.0	23.4
RU	30.6	31.9	30.4	**30.3**	**30.3**
CZ	20.8	21.8	**20.4**	**20.4**	20.8

Table 5. WER of RDLT systems: lang-dep - language-specific transformation training, EN only - transformation trained on English, 8L w/o VI - transformation trained on all 8 languages without Vietnamese, 8L - transformation trained on all 8 languages, 8L+CTS - transformation trained on all 8 languages plus CTS.

is trained, and where GMM posteriors can trigger-on transformations according to the acoustic region of the current or neighboring frames. Moreover, it can also switch-on transformation corresponding to a language-specific sound, when it is used in the multiple language scenario.

GMM was built by clustering of Gaussians from all eight language dependent systems to cover as much acoustic variability as possible. Again, HLDA preceding RDLT estimated from all languages ("HLDA 8L" in table 3) served as a starting point in all experiments to ensure fair comparison of all results.

The first two columns in table 5 report 1%-3.2% drop in accuracy if only English transformation was used (compared to RDLT trained on particular language). "8L w/o VI" presents the results of systems with transformation trained on all languages excluding Vietnamese. It is interesting to see that on Vietnamese (which was not seen during the training), such transformation performs 1.4% better than English only one, showing the advantage of multi-lingual training. The last two column "8L" and "8L+CTS" show results of transformations trained on all languages and also with added telephone speech. Here, we can see significant improvement from multilingual training - single language baselines (lang-dep) were beaten in most of the cases.

7. CONCLUSION AND FUTURE WORK

We successfully tested multilingual training of feature transformations. Statistics required by transformation update were collected by each language-specific system and merged. We presented this procedure with three popular transforms: HLDA, MPE-HLDA and RDLT.

In the most important experiment with RDLT, multilingual training brought 0.2%-0.5% absolute improvement over language-specific systems. Decrease of accuracy was found only for Vietnamese (0.4%). It was expected as this language is strongly different (tonal language) from the others.

Our preliminary results have shown that using language-specific GMM model in RDLT can improve accuracy by 0.4% absolute, however, the feature extraction can not be shared in this case. Building a more accurate language-independent GMM model is an open field for future work.

8. REFERENCES

[1] D. Povey, L. Burget, M. Agarwal, P. Akyazi, A. Ghoshal, O. Glembek, K. N. Goel, M. Karafiát, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "The subspace gaussian mixture model-a structured model for speech recognition," *Computer Speech and Language*, vol. 25, no. 2, pp. 404–439, 2011.

[2] N. Kumar, "Investigation of silicon-auditory models and generalization of linear discriminant analysis for improved speech recognition," Ph.D. dissertation, John Hopkins University, Baltimore, 1997.

[3] B. Zhang and S. Matsoukas, "Minimum phoneme error based heteroscedastic linear discriminant analysis for speech recognition," in *Proc. IEEE Int. Conf. on Acoustic, Speech and Signal processing (ICASSP)*, Philadelphia, PA, USA, march 2005, pp. 925–929.

[4] B. Zhang, S. Matsoukas, and R. Schwartz, "Recent progress on the discriminative region-dependent transform for speech feature extraction," in *Proc. of Interspeech2006*, Pittsburgh, PA, USA, Sep 2006, pp. 2977–2980.

[5] A. Stolcke, F. Grézl, M. 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.

[6] F. Grézl, M. Karafiát, and M. Janda, "Study of probabilistic and bottle-neck features in multilingual environment," in *Proc. ASRU 2011*, dec 2011.

[7] M. Gales, "Semi-tied covariance matrices for hidden markov models," *IEEE Trans. Speech and Audio Processing*, vol. 7, pp. 272–281, 1999.

[8] D. Povey, "Discriminative training for large vocabulary speech recognition," Ph.D. dissertation, University of Cambridge, 2003.

[9] D. Povey, B. Kingsbury, L. Mangu, G. Saon, H. Soltau, and G. Zweig, "fmpe: Discriminatively trained features for speech recognition," in *in Proc. IEEE ICASSP*, 2005.

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

[11] T. Schultz, M. Westphal, and A. Waibel, "The globalphone project: Multilingual lvcsr with janus-3," in *in Multilingual Information Retrieval Dialogs: 2nd SQEL Workshop, Plzen, Czech Republic*, 1997, pp. 20–27.

[12] M. Karafiát, L. Burget, P. Matějka, O. Glembek, and J. H. Černocký, "ivector-based discriminative adaptation for automatic speech recognition," in *Proc. ASRU 2011*, dec 2011.