Discriminative Training and Channel Compensation for Acoustic Language Recognition

Valiantsina Hubeika, Lukáš Burget, Pavel Matějka, Petr Schwarz

Speech@FIT, Brno University of Technology, Czech Republic
xhubei00@stud.fit.vutbr.cz, {burget|matejkap|schwarzp}@fit.vutbr.cz

Abstract

This paper describes the acoustic language recognition subsystems of Brno University of Technology (BUT) which contributed to the BUT main submission to the NIST LRE 2007. Two main techniques are employed in the subsystems discriminative training in terms of Maximum Mutual Information, and channel compensation in terms of eigenchannel adaptation in both, model and feature domain. The complementarity of the approaches is analyzed.

Index Terms: Language detection, NIST LRE 2007 evaluation, discriminative training, eigenchannel adaptation in model domain, eigenchannel adaptation in feature domain

1. Introduction

To date, there is a fair number of methods developed to improve performance of the state-of-the-art acoustic language recognition systems. Still, two issues are main challenges in the task, inter-session channel variability compensation as recordings belonging to the same language may be obtained through different channels, and language discrimination as some languages may have common features. This paper addressed both these problems within the UBM-GMM framework [12]. Here, to compensate on the channel, eigenchannel adaptation technique is applied; to train the models descriptively, Maximum Mutual Information (MMI) is used.

Formerly, a channel compensation method was proposed task by Kenny [22] in terms of factor analysis (FA). Brümmer [13] has developed a simplified version of FA, eigenchannel adaptation. These methods were developed within GMM framework and are implemented in model domain. Later, Castaldo in [7] has introduced an approximation of eigenchannel adaptation, eigenchannel adaptation in feature domain. With channel compensation performed in feature domain, different approaches can be used for the feature distribution modeling. Both compensating techniques, eigenchannel adaptation in model and feature domain, were involved in our systems.

As was proven during LRE 2005 in [2], discriminative training, by means of MMI, in language recognition task is highly beneficial and brought a great decrease in EER.

We investigate improvements given by both approaches and their combination. Further, we examine complementarity of the both methods and systems based on approaches of different nature, such as phonotactic systems.

2. Theoretical Background

This section gives a brief information on the objectives of eigenchannel adaptation and discriminative training.

2.1. Eigenchannel Adaptation in Model Domain

Let supervector be a MD dimensional vector constructed by concatenating all GMM mean vectors and normalized by corresponding standard deviations. M is the number of Gaussian mixture components in GMM and D is dimensionality of features. Before eigenchannel adaptation can be applied, we must identify directions in which supervector is mostly affected by changing channel. These directions (eigenchannels) are defined by columns of $MD \times R$ matrix $\mathbf{V}$, where R is the chosen number of eigenchannels ($R = 50$ in our system). The matrix $\mathbf{V}$ is given then by R eigenvectors of average within-class covariance matrix, where each class is represented by supervectors estimated on different segments of the same language.

Once the eigenchannels are identified, language-dependent model (or language-independent UBM) can be adapted to a test conversation by shifting its supervector in the directions given by eigenchannels to better fit the test conversation data. Mathematically, this can be expressed as finding the channel factors, x, that maximize the following MAP criterion:

$$p(\mathbf{O}|\mathbf{s} + \mathbf{V}\mathbf{x})N(\mathbf{x}; \mathbf{0}, \mathbf{I}) \tag{1}$$

where $\mathbf{s}$ is supervector representing the model to be adapted, $p(\mathbf{O}|\mathbf{s} + \mathbf{V}\mathbf{x})$ is likelihood of the test conversation given the adapted supervector (model) and $N(\mathbf{x}; \mathbf{0}, \mathbf{I})$ denotes normally distributed vector. Assuming fixed occupation of Gaussian mixture components by test conversation frames, $\mathbf{o}_t, t = 1, \ldots, T$, it can be shown [13] that $\mathbf{x}$ maximizing criterion (1) is given by:

$$\mathbf{x} = \mathbf{A}^{-1} \sum_{m=1}^{M} \mathbf{V}_m^T \sum_{t=1}^{T} \gamma_m(t) \frac{\mathbf{o}_t - \mu_m}{\sigma_m} \tag{2}$$

where $\mathbf{V}_m$ is $D \times R$ part of matrix $\mathbf{V}$ corresponding to m^{th} mixture component, $\gamma_m(t)$ is the probability of occupation mixture component m at time t, μ_m and σ_m are the mixture component's mean and standard deviation vectors of the model to be adapted and

$$\mathbf{A} = \mathbf{I} + \sum_{m=1}^{M} \mathbf{V}_m^T \mathbf{V}_m \sum_{t=1}^{T} \gamma_m(t). \tag{3}$$

In our implementation, occupation probabilities, $\gamma_m(t)$, are computed using UBM and assumed to be fixed for given test conversation.

2.2. Eigenchannel Adaptation in Feature Domain

Adaptation in feature domain aims at projecting every observation feature $\mathbf{o}(t)$ to the session-independent space. Channel factors, $\mathbf{x}$, are estimated using UBM (and not speaker-dependent

models). The adapted feature vector is then obtained using 1-best Gaussian in the following way:

$$\mathbf{o}_t^{'} = \mathbf{o}_t + \mathbf{V}_m \mathbf{x} \quad (4)$$

where m is the index of the best scored Gaussian and $\mathbf{V}_m$ is the part of $\mathbf{V}$ corresponding to the m-th Gaussian.

2.3. Maximum Mutual Training

Unlike in the case of ML training which aims to maximize the overall likelihood of training data given the transcriptions, the MMI objective function to maximize is the posterior probability of correctly recognizing all training segments:

$$F_{MMI}(\lambda) = \sum_{r=1}^{R} \log \frac{p_\lambda(O_r|s_r)^{K_r} P(S_r)}{\sum_{\forall s} p_\lambda(O_r|s)^{K_r} P(s)} \quad (5)$$

where $p_\lambda(O_r|s_r)$ is likelihood of r-th training segment, O_r, given the correct transcription of the segment, s_r, and model parameters, λ. R is the number of training segments and the denominator represents the overall probability density, $p_\lambda(O_r)$. Definition of the re-estimation formula is to be found in [2].

3. Experimental Setup

The results are presented in terms of the $100 \times C_{avg}$ (the formulas are to be found in [17]).

3.1. Data

3.1.1. Training Data

To compile the training data set, different sources were used (NIST1996, NIST2003, NIST2005, CallHome, CallFriend, Fisher, Mixer, OGI-multilingual, OGI 22 languages, Foreigen Accented Englis, SpeechDat-East) [20]. The amount of training data for different languages greatly varied, from 1.5h for Thai language to 228h for English.

The training data was divided onto two subsets: the first subset was used for training the models of languages and the second was used for training of the back-end parameters.

3.1.2. Evaluation data

NIST LRE2007 data was used as the evaluation data. There are 14 languages defined as detection targets with more than 7500 segments to identify. The evaluation set contains test segments with three nominal durations of speech: 3, 10 and 30 seconds. Detailed information can be found in the NIST LRE 2007 evaluation plan [17].

3.2. Systems

3.2.1. Pre-processing

The voice activity detection (VAD) is performed by our Hungarian phoneme recognizer [15], with all the phoneme classes linked to 'speech' class. The frames containing silence are excluded from the further processing.

3.2.2. Features

All systems use the shifted-delta-cepstra (SDC) [1] together with direct MFCC. The feature extraction was the same as in our LRE 2005 system [2]: 7 MFCC coefficients (including coefficient C0) concatenated with SDC 7-1-3-7, which totals in 56 coefficients per frame.

The features were transformed using vocal-tract length normalization (VTLN) [5]. The warping factors are estimated using single GMM (512 Gaussians), ML-trained on the whole CallFriend database (using all the languages). The model was trained in standard speaker adaptive training (SAT) fashion in four iterations of alternately re-estimating the model parameters and the warping factors for the training data.

3.2.3. GMM system with 2048 Gaussians per language with eigenchannel adaptation in model domain: `GMM2048-eigchan`

The inspiration comes from our GMM system for speaker recognition [14] which follow conventional Universal Background Model-Gaussian Mixture Modeling (UBM-GMM) paradigm [12].

Each language-dependent model is obtained by traditional *relevance MAP* adaptation [4] of UBM using enrollment conversation. Only the means are adapted with the relevance factor $\tau = 19$.

In the verification phase, standard Top-N Expected Log Likelihood Ratio (ELLR) scoring [4] is used to obtain verification score, where $N = 10$ in our system. However, for each trial, both the language-dependent model and the UBM are adapted to the channel of the test conversation using eigenchannel adaptation in model domain prior to computing the log likelihood ratio score.

The eigenchannel matrix was composed of eigenchannels derived in the following way:

1. UBM is trained using the original features.
2. For each utterance, a new GMM is obtained by MAP adaptation.
3. A supervector of means normalized by corresponding standard deviations is obtained from each GMM.
4. A maximum of 100 supervectors per database and language were selected.
5. The mean is subtracted from supervectors over each language of a database (not over language as one would expect)
6. Eigenchannels (i.e. directions in which language-dependent models are adapted for each test utterance) are given by eigen vectors of the covariance matrix estimated from the supervectors (see [3] for details).

3.2.4. GMM system with 2048 Gaussians per language with eigenchannel adaptation in feature domain `GMM2048-chcf`

A similar set of GMM models with 2048 Gaussians per language was trained in UBM-GMM fashion. However, the features (both, the training and test set) were first compensated using eigenchannel adaptation in feature domain [10, 11] (where eigenchannel matrix was the same as in the standard approach, see 3.2.3). In the case of the training data, the channel factors (see equation 1) were estimated using the UBM with 2048 Gaussians. The test data was channel compensated in the same manner as the training data. However, due to the short duration of the segments, to achieve better generalization (as eigenchannels can be estimated more robustly from the covariance matrix), the UBM with 256 Gaussians was used for channel factor estimation.

Table 1: Performance of our acoustic systems on LRE 2007 data

	30 sec	10sec	3sec
GMM2048, baseline	8.03	12.89	21.77
GMM2048-eigchan	2.76	7.38	17.14
GMM2048-chcf	2.94	7.40	17.93
GMM256-MMI (15 MMI it)	4.15	8.61	18.43
GMM256-MMI-chcf (3 MMI it)	3.73	9.81	20.98
GMM2048-MMI-chcf (3 MMI it)	2.41	7.02	16.90

Table 2: Performance of our best-performing acoustic and phonotactic system, and their fusion

	30 sec	10sec	3sec
(1) GMM2048-MMI-chcf	2.41	7.02	16.90
(2) EN_Tree	3.54	10.69	22.66
(1) + (2) (LDA fusion)	1.50	5.27	14.55

3.2.5. GMM-MMI: `GMM256-MMI`

This system uses GMM models with 256 Gaussians per language as the base models, where mean and variance parameters were iteratively re-estimated using Maximum Mutual Information criterion - the same as for LRE2005 [2]. A relatively small number of Gaussians was chosen for high resource consumption during MMI training. The models' parameters were re-estimated in 15 iterations.

3.2.6. GMM-MMI with channel compensated features: `GMM256-MMI-chcf`, `GMM2048-MMI-chcf`

The GMM256-MMI-chcf system was trained in an identical manner as the GMM256-MMI system, however the features were preliminary compensated by means of eigenchannel adaptation in feature domain.

In the GMM2048-MMI-chcf system the number of Gaussians per language was increased to 2048.

3.3. Normalization and Calibration

In this work, all results are presented for the systems calibrated using linear Gaussian back-end (LDA) and linear logistic regression back-end (LLR) [8] used in cascade. During LDA, for each class, a single full-covariance Gaussian (the covariance matrix is shared among all classes) is trained on the vector of scores generated from all models. LLR is trained in a discriminative fashion. The FoCal Multi-class toolkit by Niko Brummer[1] was used for this purpose.

4. Results

We used a UBM-GMM system with 2048 Gaussians per language as the baseline system, where no eigenchannel adaptation was employed (GMM2048). Results of the individual systems described above and the baseline are listed in Table 1.

When eigenchannel adaptation in model domain was applied, GMM2048-eigchan, the error decreased almost to one third of the baseline. When eigenchannel adaptation was done in feature domain, GMM2048-chcf, the error was slightly higher than for GMM2048-eigenchan but the approach enables simple application of additional MMI parameter re-training to improve the performance.

Then several experiments were run by applying MMI training in order to select the best performing configuration. Inspired by our 2005 LID system, GMM-MMI system was first trained with 256 Gaussians. In this case, 15 iterations of the parameter re-estimations were required to converge. the error of this system was significantly lower than the error of the baseline, however the system did not reach the performance of the GMM2048-eigchan system.

Observing the good performance of the systems employing eigenchannel adaptation and MMI training, respectively, and assuming complementarity of the techniques, our intention was to combine both techniques in order to achieve further improvement of the result. When the models with 256 Gaussians were trained on the compensated features and the parameters of the models were re-estimated by means of MMI, where already 3 iterations were sufficient, we observed relative improvements of 22 % to the accuracy of the GMM256-MMI system on 30 sec condition.

Still, we supposed there was room for further improving of the recognition by increasing the number of Gaussians. When the models were trained in the same manner as GMM256-MMI-chcf only with the number of Gaussians increased to 2048 (again, only 3 iterations were run), the system out-performed the 2048GMM-eigchan system by 35 % relative in 30 sec condition.

Table 3: Effect of calibration for the GMM2048-MMI-chcf on LRE 2007 data

	30 sec	10 sec	3 sec
No back-end	5.75	9.45	18.44
LDA+LLR	2.41	7.02	16.90

4.1. Calibration

The calibration of the obtained scores was an important part in building our systems. To outline the effect of the calibration, the results of the uncalibrated GMM2048-MMI-chcf system are present as well as of the calibrated system (see Tab 3). However, in case of 3 sec condition, the decrease of the error is only about 8 % relative, in case of 30 sec condition, we could observe more than 50 % of relative reduction of the error.

4.2. Complementarity with the Other System

In order to draw an overview of the performance of our acoustic systems, we present (for sake of comparison) results achieved with our best phonotactic system, EN_Tree (see Tab 2) [21]. The approach is based on recognizing of the phonemes using English phoneme recognizer and following language modeling (PRLM). The EN_Tree system employs binary decision tree language modeling based on creating a single language independent tree (UBM) and adapting its distributions to individual language training data, as described in Navratil's work [18, 19]. Binary decision tree is trained on posterior weighted counts from phoneme lattices [2]. When both, our best-performing acoustic system GMM2048-chcf and EN_Tree, were fused, we observed a great reduction in ERR which indicates high com-

[1] http://niko.brummer.googlepages.com/focalmulticlass

plementarity of the systems. Complementarity of our other systems was further examined, for a detailed description see [20].

5. Conclusion

We showed that both eigenchannel adaptation and MMI training are greatly beneficial in the language recognition task. It was shown that, the approximation of the standard eigenchannel adaptation, eigenchannel adaptation in feature domain is almost as accurate as the standard approach. Moreover, it has a great advantage, that it allows to apply MMI parameter re-estimation without modifying the MMI training algorithm. We showed that when eigenchannel adaptation is applied in feature domain, further improvement of the result can be achieved by subsequent re-estimating of the parameter of GMM by using MMI training. We showed that our best acoustic system is complementary and well fused with our other systems. We have also shown, the calibration of the obtained scores is an important part of building an accurate recognition system.

6. Acknowledgment

This work was partly supported by European projects AMIDA (FP6-033812), Caretaker (FP6-027231) and MOBIO (FP7-214324), by Grant Agency of Czech Republic under project No. 102/08/0707 and by Czech Ministry of Education under project No. MSM0021630528. The hardware used in this work was partially provided by CESNET under project No. 201/2006. Lukáš Burget was supported by Grant Agency of Czech Republic under project No. GP102/06/383.

7. References

[1] P.A. Torres-Carrasquillo, E. Singer, M.A. Kohler, R.J. Greene, D.A. Reynolds, and J.R. Deller Jr., "Approaches to language identification using Gaussian mixture models and shifted delta cepstral features," in *Proc. International Conferences on Spoken Language Processing (ICSLP)*, Sept. 2002, pp. 89–92.

[2] P. Matějka, L. Burget, P. Schwarz, and J. Černocký, "Brno University of Technology system for NIST 2005 Language recognition evaluation," in *IEEE Odyssey: The Speaker and Language Recognition Workshop*, San Juan, Puerto Rico, June 2006, pp. 57–64.

[3] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky: Analysis of feature extraction and channel compensation in GMM speaker recognition system, In: IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 7, 2007, pp. 1979-1986, ISSN 1558-7916.

[4] D. A. Reynolds, "Comparison of background normalization methods for text-independent speaker verification," in *Proc. Eurospeech*, Rhodes, Greece, Sept. 1997, pp. 963–966.

[5] J. Cohen, T. Kamm, and A.G. Andreou, "Vocal tract normalization in speech recognition: Compensating for systematic speaker variability," *J. Acoust. Soc. Am.*, , no. 97, pp. 2346, 1995.

[6] E. Singer, P.A. Torres-Carrasquillo, T.P. Gleason, W.M. Campbell, and D.A. Reynolds, "Acoustic,phonetic,and discriminative approaches to automatic language identification," in *Proc. Eurospeech*, Sept. 2003, pp. 1345–1348.

[7] Castaldo, F.. Colibro, D.. Dalmasso, E.. Laface, P.. Vair, C.. "Compensation of nuisance factors for speaker and language recognition", In: IEEE Transactions on Audio, Speech, and Language Processing, Vol. 15, No. 7, 2007, pp 1969–1978, ISSN 1558-7916.

[8] N. Brmmer, L. Burget, J. Cernocky, O. Glembek, F. Grezl, M. Karafiát, D. van Leeuwen, P. Matejka, P. Schwarz, and A. Strasheim, "Fusion of heterogeneous speaker recognition systems in the STBU submission for the NIST speaker recognition evaluation 2006," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 2072–2084, Sept. 2007.

[9] W. Campbell, D.E. Sturim, D.A. Reynolds, and A. Solomonoff, "Svm based speaker verification using a GMM supervector kernel and nap variability compensation," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Toulouse, France, May 2006, vol. I, pp. 97–100.

[10] V. Hubeika, L. Burget, P. Matejka, and J. Cernocky, "Channel compensation for speaker recognition," in *Proc. Joint Workshop on Multimodal Interaction and Related Machine Learning Algorithms*, June 2007.

[11] F. Castaldo, E. Dalmasso, P. Laface, D. Colibro, and C. Vair, "Language identification using acoustic models and speaker compensated cepstral-time matrices," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing (ICASSP)*, Honolulu, Hawaii, USA, Oct. 2007, vol. 4, pp. 1013–1016.

[12] D. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "Speaker verification using adapted Gaussian mixture models," *Digital Signal Processing*, vol. 10, no. 1–3, pp. 19–41, 2000.

[13] Niko Brummer, "Spescom DataVoice NIST 2004 system description," in *Proc. NIST Speaker Recognition Evaluation 2004*, Toledo, Spain, June 2004.

[14] L. Burget, P. Matejka, P. Schwarz, O. Glembek, and J. Cernocky, "Analysis of feature extraction and channel compensation in GMM speaker recognition system," *IEEE Transactions on Audio, Speech and Language Processing*, vol. 15, no. 7, pp. 1979–1986, Sept. 2007.

[15] P. Schwarz, P. Matějka, and J. Černocký, "Towards lower error rates in phoneme recognition," in *Proc. International Conference on Text, Speech and Dialogue*, Brno, Czech Republic, Sept. 2004, pp. 465–472.

[16] W. Campbell, D. Sturim, and D. Reynolds, "Support vector machines using GMM supervectors for speaker verification," *IEEE Signal Processing Letters*, vol. 13, no. 5, pp. 308–311, 2006.

[17] NIST 2007 language recognition evaluation plan (lre07) www.nist.gov/speech/tests/lang/2007/lre07evalplanv8b.pdf

[18] J. Navratil: Spoken language recognition-a step toward multilinguality in speech processing, in IEEE Trans. on Speech and Audio Processing, Vol. 9, No. 6, pp. 678-685 ISSN: 1063-6676, September 2001.

[19] J. Navratil: "Recent advances in phonotactic language recognition using binary-decision trees," in *Proc. International Conferences on Spoken Language Processing (ICSLP)*, Pittsburgh, PA, October 2006

[20] P. Matějka at al., "Brno university of technology system for nist 2007 language recognition evaluation," in submited to: *Proc. International Conferences on Spoken Language Processing (ICSLP), Brisbane, Australia*

[21] O. Glembek et al.: Advances in phonotactic language recognition, submitted to *Proc. International Conferences on Spoken Language Processing (ICSLP), Brisbane, Australia*.

[22] P. Kenny, P. Dumouchel (2004): "Experiments in speaker verification using factor analysis likelihood ratios", in *Odyssey: The Speaker and Language Recognition Workshop* , 2004 , pp. 219–226.