Odyssey 2012
The Speaker and Language Recognition Workshop
25-28 June 2012, Singapore

Speaker vectors from Subspace Gaussian Mixture Model as complementary features for Language Identification

Oldřich Plchot[1], *Martin Karafiát*[1], *Niko Brümmer*[2], *Ondřej Glembek*[1], *Pavel Matějka*[1], *Edward de Villiers*[2] *and Jan "Honza" Černocký*[1]

(1) Brno University of Technology, Speech@FIT and IT4I Center of Excellence, Czech Republic
(2) AGNITIO, South Africa

{iplchot,karafiat,glembek,matejkap,cernocky}@fit.vutbr.cz[1], {niko.brummer,edwarddsp}@gmail.com[2]

Abstract

In this paper, we explore new high-level features for language identification. The recently introduced Subspace Gaussian Mixture Models (SGMM) provide an elegant and efficient way for GMM acoustic modelling, with mean supervectors represented in a low-dimensional representative subspace. SGMMs also provide an efficient way of speaker adaptation by means of low-dimensional vectors. In our framework, these vectors are used as features for language identification. They are compared with our acoustic iVector system, which architecture is currently considered state-of-the-art for Language Identification and Speaker Verification. The results of both systems and their fusion are reported on the NIST LRE2009 dataset.

1. Introduction

Current language identification (LID) systems are divided into two broad categories: phonotactic and acoustic. In the past several years, the performance improvements of both approaches have mainly been due to subspace techniques, such as Joint Factor Analysis (JFA) [1] [2], general representative subspaces (known as iVectors) [3, 4] and subspace multinomial modelling [5].

There is however a third group of LID approaches building on high-level features. Among these, the maximum likelihood linear regression (MLLR) and constrained MLLR (CMLLR) adaptation matrices from phone recognition or large vocabulary continuous speech recognition (LVCSR) systems have proven to be useful in speaker verification and have already been tested in LID [6] with encouraging results.

Another subspace acoustic modelling technique is the Subspace Gaussian Mixture Model (SGMM) [7]. It was initially developed at the Johns Hopkins University 2009 summer workshop titled "Low Development Cost, High Quality Speech Recognition for New Languages and Domains".

This framework allows not only for the modelling itself, but also for an efficient speaker adaptation of the models using low-dimensional vectors in a "speaker subspace" [8]. We use the SGMM speaker adaptation vectors as features for LID in a similar manner to SRI's use of MLLR speaker adaptation matrices for the same purpose [6]. The cost of getting these vectors is in fact not very high, because we already run our LVCSR system to obtain other features, like counts for the phonotactic systems. We then output the SGMM adaptation vectors at the same time.

The outline of the paper is as follows: Section 2 presents the SGMM modelling and adaptation framework, and section 3 summarizes the iVector approach. Section 5 outlines the classifiers and fusion used. Sections 4 and 6 deal with data and experimental results respectively, and section 7 concludes the paper.

This work was partly supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. D10PC20015, by Czech Ministry of Trade and Commerce project No. FR-TI1/034, by Czech Ministry of Education project No. MSM0021630528 and by European Regional Development Fund in the IT4Innovations Centre of Excellence project (CZ.1.05/1.1.00/02.0070).

2. SGMM

In conventional acoustic models for speech recognition, the distribution of each (possibly tied) HMM state is represented by a relatively large number of parameters completely defining a Gaussian Mixture Model (GMM). Although an SGMM also assumes mixtures of Gaussians to be the underlying state distribution, the high-dimensional supervector of all GMM parameters is constrained to live in a relatively low-dimensional subspace, which is common to all states. This constraint is justified by a high correlation between distributions of states (especially in the case of many similar context-dependent models) and by realizing that the variety of distributions corresponding to the sounds producible by the human articulatory tract is quite limited. The majority of the parameters in an SGMM are the parameters shared across the states defining the subspace of the possible GMM parameters. Distributions of the individual states are then described using relatively low-dimensional vectors representing coordinates in such a subspace. Therefore, an SGMM allows for a much more compact representation of HMM state distributions, which results in more robust estimation of parameters and improved performance especially when a limited amount of training data is available.

When an SGMM is used as the acoustic model for speech recognition, the distribution of features for HMM state j is modelled as a mixture of Gaussian distributions:

$$p(\mathbf{x}|j) = \sum_{i=1}^{I} w_{ji} \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}_{ji}, \boldsymbol{\Sigma}_i), \qquad (1)$$

where the same number of mixture components I (typically few hundred for SGMMs) is used for all states. Across the states, the corresponding mixture components share the same full covariance matrix $\boldsymbol{\Sigma}_i$. Unlike in a conventional GMM model, the mean vectors $\boldsymbol{\mu}_{ji}$ and mixture weights w_{ji} are not directly estimated as parameters of the model. Instead, for a particular state

j, all mean vectors are represented by a single low-dimensional vector $\mathbf{v}_j$ (typically few tens of coefficients), from which the mean vectors are derived as

$$\boldsymbol{\mu}_{ji} = \mathbf{M}_i \mathbf{v}_j. \tag{2}$$

We expand the model by speaker adaptation, where we use a Constrained MLLR (CMLLR) paradigm of feature transformation [9]. We replace our feature $\mathbf{x}$ with the transformed feature

$$\mathbf{x}' = \mathbf{A}^{(s)}\mathbf{x} + \mathbf{b}^{(s)}, \tag{3}$$

where s is the speaker index, $\mathbf{A}^{(s)}$ is speaker-dependent adaptation matrix and $\mathbf{b}^{(s)}$ is a speaker-dependent offset. In the SGMM framework, we apply the CMLLR technique by adding a speaker-dependent offset for each Gaussian index i in the form of a term $\mathbf{N}_i\mathbf{v}^{(s)}$. Equation (2) for the mean vectors now becomes:

$$\boldsymbol{\mu}_{ji} = \mathbf{M}_i \mathbf{v}_j + \mathbf{N}_i \mathbf{v}^{(s)}, \tag{4}$$

where $\mathbf{v}^{(s)}$ is a "speaker vector", defining the transform in "speaker subspace" $\mathbf{N}_i$. Vectors $\mathbf{v}^{(s)}$ estimated for every segment are used as features for training the LID classifier. For the complete set of formulae and details on SGMMs, see [10, 8].

3. Acoustic iVector System

Let us briefly recall the definition of iVectors. Note that the formulae were derived for speaker recognition in [11] and iVectors have already been used for LID [3, 4]. The main idea is that the language- and channel-dependent GMM supervector $\mathbf{l}$ can be modelled as:

$$\mathbf{l} = \mathbf{m} + \mathbf{T}\mathbf{v}, \tag{5}$$

where $\mathbf{m}$ is the UBM GMM mean supervector, $\mathbf{T}$ is a low-rank matrix representing M bases spanning a subspace with important variability in the mean supervector space, and $\mathbf{v}$ is a standard normal distributed vector of size M.

For each observation $\mathcal{X}$, the aim is to estimate the parameters of the posterior probability of $\mathbf{v}$:

$$\mathrm{p}(\mathbf{v}|\mathcal{X}) = \mathcal{N}(\mathbf{v}; \mathbf{v}_\mathcal{X}, \mathbf{L}_\mathcal{X}^{-1}). \tag{6}$$

The iVector is the MAP point estimate of the variable $\mathbf{v}$, i.e. the mean $\mathbf{v}_\mathcal{X}$ of the posterior distribution $\mathrm{p}(\mathbf{v}|\mathcal{X})$. It maps most of the relevant information from a variable-length observation $\mathcal{X}$ to a fixed- (small-) dimensional vector. $\mathbf{T}$ is referred to as the iVector extractor.

4. Data

The following data (distributed by the LDC and ELRA) were used to train our systems:

CallFriend
Fisher English Part 1. and 2.
Fisher Levantine Arabic
HKUST Mandarin
Mixer (data from NIST SRE 2004, 2005, 2006, 2008)
development data for NIST LRE 2007
OGI-multilingual
OGI 22 languages
Foreign Accented English
SpeechDat-East
SwitchBoard
Voice of America radio broadcast

4.1. Training and development data

Our data was split into two independent subsets, denoted TRAIN and DEV. The TRAIN subset contained the 23 target languages of NIST LRE2009 and had about 50 000 segments in total. The DEV subset also contained only the 23 target languages in about 38 000 segments. The DEV subset was split into balanced subsets having nominal durations of 3s, 10s and 30s. The DEV set was based on segments from previous evaluations plus additional segments extracted from longer files from CTS, VOA3 and human-audited VOA2 data, which were not contained in the TRAIN set. A more detailed overview of the composition and processing of our DEV set is given in [12].

4.2. Test data — NIST LRE 2009

Our test data was the NIST LRE2009 evaluation dataset, which contains 23 target languages and 16 out-of-set languages. It contains mainly telephone call segments coming from the VOA broadcasts and for some languages it also contains heldout conversational telephone speech segments collected for the previous evaluations.[1] The process of collecting the VOA database is described in [13].

5. Recognizer training and fusion

We used three different discriminative training stages to build a fused language recognizer. The first, *basic recognizer* stage, performs dimensionality reduction from iVectors to scores and was trained on the TRAIN subset of the data. The other two stages do *precalibration* and *fusion* in score-space and are both trained on the DEV subset. All stages are implemented with different flavors of multiclass logistic regression [14].

5.1. Basic recognizer

Two similar logistic regression recognizers were independently trained (on the TRAIN subset) for the two different feature vectors, namely acoustic and SGMM iVectors. Each recognizer uses an affine transform to convert the D-dimensional iVector, $\mathbf{v}_t$, for trial t, into a K-dimensional score-vector, $\mathbf{s}_t$:

$$\mathbf{s}_t = \mathbf{A}\mathbf{T}\mathbf{v}_t + \mathbf{b}, \tag{7}$$

where $K = 23$ is the number of target languages. $\mathbf{T}$ is a D-by-D matrix which performs within-class covariance normalization (WCCN), such that the mean class-conditional sample covariance matrix over the training data becomes identity. The logistic regression parameters are $\mathbf{A}$, a K-by-D matrix, and $\mathbf{b}$, a K-dimensional vector and they are trained by minimizing the regularized objective function:

$$\lambda\,\mathrm{tr}(\mathbf{A}^\mathrm{T}\mathbf{A}) - \sum_{i=1}^{K} \frac{1}{KN_i} \sum_{t\in\mathcal{S}_i} \log \frac{\exp(s_{it})}{\sum_{j=1}^{K}\exp(s_{jt})}, \tag{8}$$

where s_{it} is the ith component of $\mathbf{s}_t$ and $\mathcal{S}_i$ is the set of N_i training examples of language i. The regularization weight, λ, was set to:

$$\lambda = \frac{1}{N}\left(\frac{1}{N}\sum_{t\in\mathcal{S}}\sqrt{\mathbf{v}_t^\mathrm{T}\mathbf{T}^\mathrm{T}\mathbf{T}\mathbf{v}_t}\right)^2, \tag{9}$$

[1] Detailed distribution of segments per language can be found on the official NIST LRE09 results page — `http://www.itl.nist.gov/iad/mig/tests/lre/2009/lre09_eval_results/index.html`

where $N = \sum_i N_i$ and $\mathcal{S} = \bigcup_i \mathcal{S}_i$. We refer to the two terms of (8) as the *regularization penalty* and the *multiclass cross-entropy* and we note that they were almost equal at the minimum, suggesting that regularization plays an important role.

5.2. Precalibration

Each of the two recognizers was independently precalibrated with an affine transform trained on the independent DEV dataset:

$$\mathbf{r}_t = \mathbf{C}\mathbf{s}_t + \mathbf{d} \tag{10}$$

where $\mathbf{C}$ is a full K-by-K matrix and $\mathbf{d}$ is a K-dimensional vector. Note that the pre-calibration does not change the score-vector dimensionality. These parameters were again trained by regularized logistic regression, but here, no WCCN were applied. For the SGMM system, some trials failed to produce speaker vectors. After pre-calibration, we inserted zero vectors, $\mathbf{r}_t = \mathbf{0}$, for such trials.

5.3. Fusion

Let $\mathbf{r}_{t1}$ and $\mathbf{r}_{t2}$ denote the outputs of the precalibrated acoustic and SGMM recognizers. These outputs were fused as:

$$\boldsymbol{\ell}_t = \alpha_1 \mathbf{r}_{t1} + \alpha_2 \mathbf{r}_{t2} + \boldsymbol{\beta} \tag{11}$$

where α_1, α_2 are scalar weights and $\boldsymbol{\beta}$ is a K-dimensional vector. These parameters were again trained by multiclass logistic regression on the DEV subset. Here neither WCCN, nor regularization, was applied.

After fusion, the 23 components of $\boldsymbol{\ell}$ were used as language log-likelihoods and were used as is to make minimum-expected-cost Bayes decisions for the cost function as prescribed by the LRE'09 evaluation plan.

6. Experiments

6.1. iVector Front-End

The acoustic system is based on MFCC/SDC acoustic features. We give a brief summary of our acoustic feature extraction and UBM training. For more detail, see our previous work [15, 16].

The inputs to the language recognizer are segments of recorded speech of varying duration. The voice activity detection (VAD) is performed by our Hungarian phoneme recognizer, with all the phoneme classes linked to the 'speech' class.

All acoustic systems used the popular shifted-delta-cepstras (SDCs) [17]. The feature extraction is similar to the BUT LRE 2005 system [16]. Every speech segment is mapped to a variable-length sequence of feature vectors as follows: After discarding silence portions, every 10 ms speech-frame is mapped to a 56-dimensional feature vector. The feature vector is the concatenation of an SDC-7-1-3-7 vector and 7 MFCC coefficients (including C0). VTLN, Cepstral mean and variance normalization and RASTA filtering are applied before SDC.

Vocal-tract length normalization (VTLN) performs simple speaker adaptation. We used MAP adaptation from UBM (single GMM with 32 diagonal Gaussians trained on Switchboard) to derive specific models for each warping factor [18]. Models are retrained using MMI (Maximum Mutual Information) criterion. The reference warping factors are obtained from the LVCSR system. The models are trained only on English data.

A 2048-component, language-independent, maximum-likelihood GMM was trained with the EM-algorithm on the pooled acoustic feature vectors of all 23 languages in the TRAIN data-set. We follow speaker recognition terminology and refer to this language-independent GMM as the *universal background model*, or UBM [19].

We used a full covariance UBM to generate zero and first order statistics, which were subsequently used for iVector extractor training. The output was a 400-dimensional vector for every speech segment.

Table 1: Cavg in % for three conditions of NIST LRE2009

System/Condition	30s	10s	3s
iVector	2.35	4.91	14.04
SGMM	12.42	16.68	28.85
Fusion	1.78	4.13	13.77

6.2. SGMM Front-end

The LVCSR system is based on cross-word tied-state triphones, with approximately 8000 tied states. It was trained on 270 hours of US English (Switchboard and Callhome corpora). The features were 13 Mel-Frequency VTLN PLP coefficients, augmented with their deltas, double-deltas and triple-deltas. VTLN used the same estimation of warping factors as described in section 6.1. Cepstral mean and variance normalization was applied with the mean and variance vectors estimated on each conversation side.

SGMM models were first initialized as standard GMMs. We used an SGMM configuration with $I = 400$ mixture components and 40-dimensional state-specific vector $\mathbf{v}_j$ and 50 000 states. This system was extended by a speaker specific part $\mathbf{N}_i \mathbf{v}^{(s)}$. The speaker-specific vectors were 400-dimensional.

In our experiments, we did not use full LVCSR with a language model. Instead, every utterance was processed by a phone recognizer to get a sequence of models, on which the vector $\mathbf{v}^{(s)}$ was estimated. These vectors were further used as features for LID.

6.3. Results

The performance of the two individual systems and their fusion is evaluated on the NIST LRE2009 closed set condition. In table 1, we report the classical metric — Average Detection Cost (Cavg) — as defined by NIST [20]. In table 2, we also report Average Pair Error-rate (PER) and Average Pair Error-rate for the N worst pairs (PER_N), where N is the number of target languages (PER_23). PER_N is the overall performance measure set by NIST for the LRE2011 evaluation [21].

Table 2: Pair error rates in % for the three conditions

	PER			PER_23		
	30s	10s	3s	30s	10s	3s
iVector	0.73	2.08	8.78	5.15	9.15	18.95
SGMM	5.48	10.21	21.72	17.68	22.95	31.99
Fusion	0.71	1.94	8.01	5.38	9.08	18.10

6.4. Discussion

The results in table 1, where we measure the ability of the system to discriminate between N languages, show that the new speaker adaptation vectors from the SGMM model contain complementary information to the acoustic system.

On the other hand, when we measure the ability of the system (given the speech segment) to decide whether the segment belongs to language L1 or L2, the contribution of the SGMM features is much smaller. The results presented in table 2 indicate that if we concentrate only on the N most difficult trials, the gain from the new features is negligible, and for the 30s condition, SGMMs even cause the iVector results to deteriorate. The belief is that SGMMs capture the speaker-specific information, and for difficult (closed) language pairs, the speaker populations can be very similar, even overlapping. SGMMs therefore obtain very similar information and thus do not bring substantial improvement over the baseline system.

7. Conclusions

We have presented new features for LID, based on speaker adaptation vectors from sub-space GMMs. We have demonstrated, that although their performance is worse than a state-of-the-art iVector system, they provide complementary information useful in the fusion. For difficult language pairs (on which the current NIST evaluations are focused), their contribution is however small and they seem to degrade the results for long utterances. We should note that, contrary to previous work making use of adaptation matrices in LID [6], our adaptation scheme was fairly simple and did not distinguish between acoustic classes. Our future work will follow this avenue, and investigate current developments in the field of SGMM modelling.

8. References

[1] N. Brümmer, A. Strassheim, V. Hubeika, P. Matějka, L. Burget, and O. Glembek, "Discriminative Acoustic Language Recognition via Channel-Compensated GMM Statistics," in *Proc. of Interspeech 2009*.

[2] N. Dehak, P. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-End Factor Analysis For Speaker Verification," *Audio, Speech, and Language Processing, IEEE Transactions on*, vol. 19, no. 4, 2010.

[3] D. Martínez, O. Plchot, L. Burget, O. Glembek, and P. Matějka, "Language Recognition in iVectors Space," in *Proc. of Interspeech 2011*.

[4] N. Dehak, P. A. Torres-Carrasquillo, D. Reynolds, and R. Dehak, "Language Recognition via Ivectors and Dimensionality Reduction," in *Proc. of Interspeech 2011*.

[5] M. Soufifar, M. Kockmann, L. Burget, O. Plchot, O. Glembek, and T. Svendsen, "iVector Approach to Phonotactic Language Recognition," in *Proc of Interspeech 2011*.

[6] A. Stolcke, M. Akbacak, L. Ferrer, S. Karajarekar, C. Richey, N. Scheffer, and E. Schriberg, "Improving Language Recognition with Multilingual Phone Recognition and Speaker Adaptation Tranforms," in *Proc. Odyssey 2010*.

[7] D. Povey, "A Tutorial Introduction to Subspace Gaussian Mixture Models for Speech Recognition," in *Tech. Rep. MSR-TR-2009-111, Microsoft Research, 2009*.

[8] D. Povey, L. Burget, M. Agarwal, P. Akyazi, K. Feng, A. Ghoshal, O. Glembek, N. K. Goel, M. Karafiát, A. Rastrow, R. Rose, P. Schwarz, and S. Thomas, "Subspace Gaussian mixture models for speech recognition," in *Proc. of ICASSP*, 2010.

[9] M. Gales, "Maximum Likelihood Linear Transformations for HMM-based Speech Recognition, Tech. Report, CUED/FINFENG/TR291, Cambridge Univ.," Tech. Rep., 1997.

[10] D. Povey, L. Burget, M. Agarwal, P. Akyazi, A. Ghoshal, O. Glembek, K. Nagendra 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.

[11] O. Glembek, L. Burget, P. Kenny, M. Karafiát, and P. Matějka, "Simplification and optimization of I-Vector Extraction," in *Proc. ICASSP 2011*, 2011, pp. 4516–4519.

[12] Z. Jančík, O. Plchot, N. Brümmer, L. Burget, O. Glembek, V. Hubeika, , M. Karafiát, P. Matějka, T. Mikolov, A. Strasheim, and Jan "Honza" Černocký, "Data selection and calibration issues in automatic language recognition – investigation with BUT-AGNITIO NIST LRE2009 system," in *Proc. Odyssey 2010*.

[13] C. Cieri et al., "The Broadcast Narrow Band Speech Corpus: a New Resource Type for Large Scale Language Recognition," in *Proc. of Interspeech 2009*.

[14] Christopher M. Bishop, *Pattern Recognition and Machine Learning (Information Science and Statistics)*, Springer, 2007.

[15] P. Matějka V. Hubeika, L. Burget and P. Schwarz, "Discriminative Training and Channel Compensation for Acoustic Language Recognition," in *Proc. Interspeech*, 2008.

[16] P. Matějka, L. Burget, P. Schwarz, and J. Černocký, "Brno University of Technology System for NIST 2005 Language Recognition Evaluation," in *Proc. NIST LRE 2005 Workshop*, San Juan, Puerto Rico, June 2006, pp. 57–64.

[17] P. A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. J. Greene, D. A. Reynolds, and J.R. Deller, "Approaches to Language Identification Using Gaussian Mixture Models and Shifted Delta Cepstral Features," in *Proc. of ICASSP 2002*.

[18] L. Welling, S. Kanthak, and H. Ney, "Improved methods for vocal tract normalization," in *Proc. ICASSP 1999*.

[19] 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, Jan. 2000.

[20] "The 2009 NIST Language Recognition Evaluation Plan (LRE09)," http://www.itl.nist.gov/iad/mig/tests/lre/2009/LRE09_EvalPlan_v6.pdf.

[21] "The 2011 NIST Language Recognition Evaluation Plan (LRE11)," http://www.nist.gov/itl/iad/mig/upload/LRE11_EvalPlan_releasev1.pdf.