Acoustic Language Identification Using Fast Discriminative Training

Fabio Castaldo [1], *Daniele Colibro*[2], *Emanuele Dalmasso*[1], *Pietro Laface*[1], *Claudio Vair*[2]

[1] Dipartimento di Automatica e Informatica, Politecnico di Torino, Italy
[2] Loquendo S.p.A, Torino, Italy
{Fabio.Castaldo,Emanuele.Dalmasso,Pietro.Laface}@polito.it
{Daniele.Colibro, Claudio.Vair}@loquendo.com

Abstract

Gaussian Mixture Models (GMMs) in combination with Support Vector Machine (SVM) classifiers have been shown to give excellent classification accuracy in speaker recognition.

In this work we use this approach for language identification, and we compare its performance with the standard approach based on GMMs.

In the GMM-SVM framework, a GMM is trained for each training or test utterance. Since it is difficult to accurately train a model with short utterances, in these conditions the standard GMMs perform better than the GMM-SVM models.

To overcome this limitation, we present an extremely fast GMM discriminative training procedure that exploits the information given by the separation hyperplanes estimated by an SVM classifier. We show that our discriminative GMMs provide considerable improvement compared with the standard GMMs and perform better than the GMM-SVM approach for short utterances, achieving state of the art performance for acoustic only systems.

Index Terms: language identification, discriminative training, GMM, SVM, separation hyperplane

1. Introduction

This paper focuses on the acoustic component of a Language Identification (LID) system. The GMM and the SVM are the state of the art classifiers [1],[2] for acoustic LID. Discriminative training of acoustic GMMs [3],[4], obtained through Maximum Mutual Information Estimation (MMIE), was demonstrated to be successful for language identification in the last formal NIST Language Recognition Evaluations (LRE) [5]. Since MMIE training requires considerable computational resources, in this work we propose a new discriminative training technique. In particular, we applied to language identification a recently proposed approach for *speaker recognition* combining Gaussian Mixture Models (GMMs) with a Support Vector Machine (SVM) classifier [6]. The results, reported in Section 4.3, which compare the performance of the GMM-SVM models with the standard GMM technique on the NIST LRE suite of recent years, clearly show the advantage of the SVM models for the 30 sec duration tests. For the short duration tests, on the other hand, such an advantage is not observed. The reason is that in the GMM-SVM framework, a GMM is trained for each *test* utterance. Thus, the duration of an utterance has a direct impact on the quality of the resulting model and on the overall LID accuracy. The problem does not exist in training because the training corpora usually include long conversations that allow robust models to be estimated.

To overcome this weakness of the GMM-SVM models, without loosing the advantages of this approach, our new discriminative training procedure for the GMMs exploits the information given by the separating hyperplanes estimated by the SVM classifiers. In particular, as will be detailed in Section 5, we shift the Gaussian means along the directions orthogonal to the hyperplane that separate each language GMM from its competitors in the space of the SVM classifier. This space is defined by a distance metric based on the approximate Kullback-Leibler (KL) divergence between GMMs. As expected, these discriminatively trained GMMs perform far better than the original models, and better than the GMM-SVM models on short duration tests.

The procedure is very fast because the GMM-SVM approach does not perform onerous iterations on all the *frames* of the training database, as required in the GMM discriminative training approaches, such as MMIE or Minimum Classification Error estimation.

The paper is organized as follows: Section 2 presents our baseline acoustic LID models, and the test databases. In Section 3 we detail the features and the database that are used to train our baseline GMMs. Section 4 summarizes the approach combining GMMs and SVM classifiers. In Section 5 we introduce our novel discriminative training procedure. Our final remarks are given in Section 6.

2. Acoustic LID models

Gaussian Mixture Models used in combination with Maximum A Posteriori (MAP) adaptation represent the core technology of most state of the art text-independent *speaker recognition* systems [1]. In these systems, the speaker models are estimated, by means of MAP adaptation, from a common GMM root model, the so-called world model or Universal Background Model (UBM). Usually, only mean vector adaptation is performed during model training. Thus, a speaker is represented by the set of mean vectors of all the Gaussians of the UBM, adapted using the speaker training data, and shares with the other speaker models the remaining UBM parameters.

MAP adaptation is not necessary in *language recognition* because every language GMM can be robustly trained by Maximum Likelihood estimation. However, we perform MAP estimation from a UBM also in LID, with a small relevance factor, for three main reasons. Language models deriving from a common UBM are required by our GMM-SVM approach. Our frame based inter-speaker variation compensation approach [8] computes its speaker factors using the UBM. A side benefit of this choice is that it allows fast selection of the Gaussians both in training and in testing. Thus, larger models can be trained discriminatively.

In the experiments described in this paper, the UBM and the language GMMs consist of mixtures of 512 Gaussians. The observation vector includes 56 parameters: the first 7

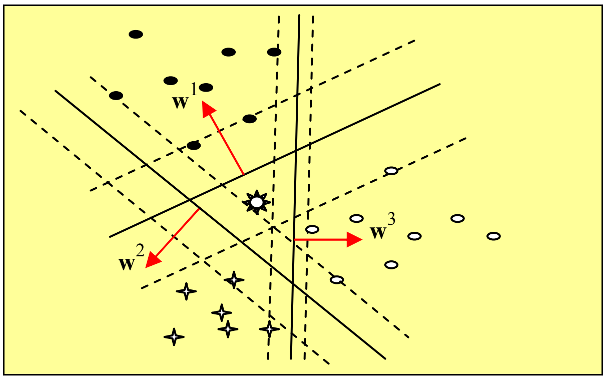

Figure 1: *Hyperplanes separating a class from the others, and their discriminative direction vectors* $\mathbf{w}^k$.

Mel frequency cepstral coefficients and their usual 7-1-3-7 Shifted Delta (SDC) features [7].

The experiments have been performed on the NIST 1996, 2003, and 2005 LRE data according to NIST evaluation rules [5]. The first two test corpora include 12 target languages: American English, Arabic, Canadian French, Farsi, German, Hindi, Japanese, Korean, Mandarin, Spanish, Tamil, and Vietnamese. Russian has been used as the out-of-language in the 2003 tests. In these evaluations there are three duration settings: 3, 10, and 30 seconds. The 1996 evaluation database consists of 1503, 1501 and 1492 sessions of 3, 10, and 30 seconds, respectively. The 2003 evaluation has 1280 trials for each duration setting. The LRE-05 corpus includes seven languages and two dialects: English-American, English-Indian, Hindi, Japanese, Korean, Mandarin-Mainland, Mandarin-Taiwan, Spanish, and Tamil. The evaluation data consists of 3662 trials for each duration setting.

3. Speaker compensated GMMs

To reduce inter-speaker variability within the same language we have shown in [8] that significant performance improvement in LID can be obtained using factor analysis. We estimate an inter-speaker subspace that represents the distortions due to inter-speaker variability, and compensate these distortions in the domain of the features. The details of this approach are given in [8] and [9].

Using compensated features, we trained a gender-dependent model for each of the 12 target languages in the NIST corpora using the training and development sets of the CallFriend [10] corpus. The conversations in this corpus were split into 8172 slices of approximately 150s. The same data sets were used for training all other types of models.

During testing, the UBM gender model that produces the best likelihood for the current utterance is selected, together with the set of its corresponding gender-dependent GMM language models. The final score for each language includes T-normalization, computed on the alternative language GMMs, followed by the log-likelihood normalization [11]:

$$\tilde{s}_l = \log\left(\frac{1}{L-1}\cdot\frac{e^{s_l}}{\sum_{k\neq l} e^{s_l}}\right) \qquad l = 1,\cdots,L \qquad (1)$$

where l and s_l are the index and the log-likelihood score of the *l-th* language GMM respectively. Gender dependent models have been used for training MMI models, while our new discriminative approach has been tested with gender independent models.

The EER reported results are the average of the EERs for each language.

4. SVM using GMM supervectors

Since Gaussian Mixture Models in combination with a Support Vector Machine classifier have been shown to give excellent classification accuracy in *speaker recognition* [6], in this work we use this approach for LID, and we compare its performance with the standard GMM based technique.

A short overview of the GMM-SVM framework is given here, focusing on the main topics that are of interest for the development of our discriminative training approach detailed in Section 5.

4.1. Linear Support Vector Machines

A linear Support Vector Machine is a two-class classifier trained to find the hyperplane which separates, with the largest margin, the samples of one class from the samples of another class. Given a set of linearly separable, labeled train data $\{\mathbf{x}_i, y_i\}$, where y_i is +1 and -1 for the positive and negative class targets respectively, the points $\mathbf{x}$ that lie on the separating hyperplane satisfy the equation

$$\mathbf{w}\cdot\mathbf{x} + b = 0 \qquad (2)$$

where $\mathbf{w}$ is the discrimination vector, which is normal to the hyperplane, $|b|/||\mathbf{w}||$ is the distance from the hyperplane to the origin, and $||\mathbf{w}||$ is the Euclidean norm of $\mathbf{w}$.

Figure 1 shows an example of hyperplanes separating a class from two other classes, and their discriminative direction vectors $\mathbf{w}^k$.

4.2. GMM supervectors

Gender independent GMMs were trained by MAP adaptation, with relevance factor 1, from a common UBM

$$f(\mathbf{x}) = \sum_g^G \omega_g \cdot \mathrm{N}\left(\mathbf{x};\boldsymbol{\mu}_g,\boldsymbol{\sigma}_g\right) \qquad (3)$$

where $N\left(\mathbf{x};\boldsymbol{\mu}_g,\boldsymbol{\sigma}_g\right)$ is a Gaussian with mean and diagonal covariance p-dimensional vectors $\boldsymbol{\mu}_g$ and $\boldsymbol{\sigma}_g$ respectively, and ω_g is its mixture weight. A specific GMM is trained for each *utterance*, both in training and in testing. Since MAP adaptation is performed only on the mean vectors, the set of the mixture weights ω_g and the diagonal covariance vectors $\boldsymbol{\sigma}_g$ are shared among all the GMMs, including the UBM.

A *pxG* supervector that maps an utterance to a high dimensional space is obtained by appending the adapted mean value of all the Gaussians of a GMM in a single stream. This mapping, however, is inaccurate because it does not take into account the weights and covariances of the Gaussians in the mixture. A more accurate mapping is obtained if the resulting supervectors can be compared according to a meaningful distance measure. The natural choice for a distance measure between two GMMs, i and j, is the approximate Kullback-Leibler divergence [12],[13], [6]:

$$D(i,j) = \sum_g^G \omega_g \sum_p \left(\frac{\mu_{gp}^i - \mu_{gp}^j}{\sigma_{gp}}\right)^2 \qquad (4)$$

where g is the *g-th* Gaussian of the mixture, and p is the dimension of the acoustic feature vector. Normalizing each component of a supervector k according to:

Table 1. *Average EER(%) of standard 512 GMMs, GMM-SVM, Discriminative GMMs, and MMI estimated GMMs classifiers on the NIST LRE tasks.*

Year	Models	Duration		
		3s	10s	30s
1996	Standard GMM	18.35	7.99	3.17
	GMM-SVM	22.27	8.29	1.41
	Discriminative GMM	16.67	5.96	1.94
	MMI GMM	14.93	4.83	1.79
2003	Standard GMM	18.60	8.75	3.84
	GMM-SVM	23.79	8.51	2.32
	Discriminative GMM	17.40	7.15	2.39
	MMI GMM	15.28	5.77	2.71
2005	Standard GMM	22.50	14.06	9.34
	GMM-SVM	25.89	14.08	6.69
	Discriminative GMM	21.43	11.80	6.96
	MMI GMM	19.16	11.76	7.79

$$\tilde{\mu}^{k}_{gp} = \sqrt{\omega_g} \cdot \frac{\mu^{k}_{gp} - \mu^{UBM}_{gp}}{\sigma_{gp}} \qquad (5)$$

the normalized UBM supervector defines the origin of a new space, where the KL divergence is a Euclidean distance.

In this high dimensional space, referred to in this paper as the KL space, an utterance model is a point whose coordinates are the supervector's parameters. The points in Figure 1 could represent utterance supervectors of different languages. The "sun" symbol, corresponding to the UBM supervector, marks the origin of this *p*x*G*-dimensional space. Since a translation does not alter the relative position, or the distance between these points, the supervector normalization term can be simply reduced to the scaling factor $\sqrt{\omega_g}/\sigma_{gp}$.

4.3. GMM–SVM

The normalized supervectors are used as samples for training a linear SVM, which produces the discrimination vectors **w** and the offset *b* in (2). The results of this approach are shown in the rows of Table 1 labeled GMM-SVM. Compared with the standard GMM classifier, the GMM-SVM system obtains far better results for the 30s duration tests. For shorter durations, however, the estimation of the utterance GMMs is not robust enough, due to the lack of data compared with the number of parameters of the GMMs. Thus, in these conditions, the GMM system gives better results.

5. GMM discriminative training

To produce more discriminative GMMs, without performing the expensive MMI estimation training, we present in this Section a new GMM training approach that exploits the information given by the separating hyperplanes estimated by the GMM-SVM classifier

Since a SVM classifier produces a discrimination vector $\mathbf{w}^k$ for each language *k*, we shift the supervector of the standard GMM of language *k* along its discriminative direction in the KL space according to

$$\hat{\boldsymbol{\mu}}^{k}\left(\alpha^{k}\right) = \tilde{\boldsymbol{\mu}}^{k} + \alpha^{k} \cdot \bar{\mathbf{w}}^{k} \qquad (6)$$

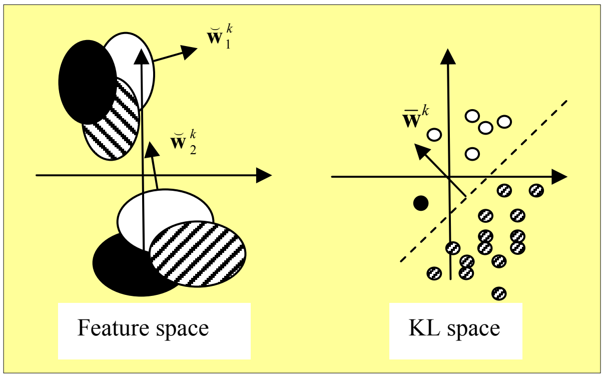

Figure 2. *Example of the original space of the acoustic features (left) and of the KL space (right). The white ellipses represent two Gaussians of language k.* $\breve{\mathbf{w}}_1^k$ *and* $\breve{\mathbf{w}}_2^k$ *, are the discriminative directions in the acoustic feature space.*

where $\bar{\mathbf{w}}^k$ is the normal to the hyperplane separating the utterance supervectors of language *k* from the supervectors of the other languages, and α^k is the shift size, which has to be found. Since the supervectors in the KL space are scaled versions of the supervectors in the original feature space, where the Gaussians have been estimated, each component of the standard GMM of language *k* will be updated according to

$$\breve{\mu}^{k}_{gp}\left(\alpha^{k}\right) = \frac{\sigma_{gp}}{\sqrt{\omega_g}} \hat{\mu}^{k}_{gp}\left(\alpha^{k}\right) \qquad (7)$$

We refer to these new models as *Discriminative GMMs*.

Figure 2 shows, on its left side, a two-dimensional acoustic feature space. The black ellipses represent two Gaussians of the UBM. The white and the dashed ellipses represent the corresponding Gaussians of two languages, the white ones referring to language *k*. The white circles shown on the right side of Figure 2 represent a two-dimensional projection of a set of utterance supervectors of language *k* mapped to the KL space. The dashed circles correspond to the utterance supervectors of the competitor languages, and the black circle is the UBM. $\breve{\mathbf{W}}_1^k$ and $\breve{\mathbf{W}}_2^k$, in the acoustic feature space, are the rescaled components of supervector $\bar{\mathbf{W}}^k$ for the two Gaussians of the language *k* GMM shown in the figure. The figure suggests that the Gaussians of a language *k* are moved away from the corresponding Gaussians of the other languages along different directions. These directions are the ones that optimize the discrimination of that language in the KL space, i.e. the directions that maximize the distance of the GMM of language *k* form its competitor GMMs. This distance increases with larger α^k, but at the same time the likelihood of each training utterance of language *k* decreases because its discriminative GMM moves away from the original MAP adapted model (which best matches the training data). This behavior is shown by the first curve in Figure 3 for a subset of 1000 utterances of the CallFriend *test* database, which has been selected as our development corpus. It shows how, for this set, the average log-likelihood ratio between the correct model and the UBM decreases as a function of α^k.

Since we cannot select largely different values for the parameters α^k, to avoid favouring the language models nearer to their original GMM, a unique parameter α will control the

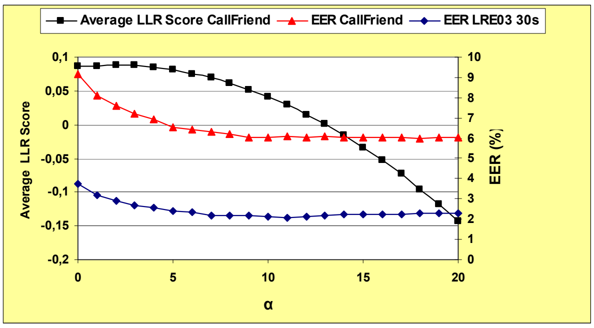

Figure 3. *Average log likelihood for a subset of 1000 utterances in the CallFriend test database, %EER for the same dataset and %EER for the 30s tests of the NIST LRE03, as a function of α.*

Table 2. *EER(%) of Discriminative GMMs, and GMM-SVM on the NIST LRE tasks. In parentheses, the average of the EERs for each language.*

Year	Models		
	Discriminative GMMs		GMM-SVM
	3s	10s	30s
1996	11.71 (13.71)	3.62 (4.92)	1.01 (1.37)
2003	13.56 (14.40)	5.50 (6.02)	1.42 (1.64)
2005	16.94 (17.85)	9.73 (11.07)	4.67 (5.81)

shift size along the discriminative directions of every language. The rule for selecting the value of α is very simple: avoid that a Gaussian move too far from the corresponding Gaussian of the UBM, possibly moving toward a region in the feature space modeled by another Gaussian, as suggested by the direction of $\breve{\mathbf{w}}_2^k$ in Figure 2. This is obtained by setting α to a value that entails maximum distance between models, but also guarantees that the average probability of the correct language model of the utterances in the development set is not less than the probability obtained by the UBM.

The validity of this criterion can be confirmed by looking at Figure 3 which shows, as a function of α, the EER obtained by the discriminative GMMs on the utterances of CallFriend development corpus and on the 30s trials of NIST LRE03. It can be noticed that low EERs are obtained, for both tests, in a range of values for α near to the zero-crossing point of the average log-likelihood ratio curve. We kept α=12 fixed for all the experiments with 512 Gaussian GMMs.

Using this simple and fast procedure for the selection of α, the discriminative GMMs provide considerable improvement compared with the standard GMMs and perform better than the GMM-SVM approach for short utterances (see Table 1).

Our procedure is much faster than MMIE training, which requires several (~20) iterations on all the *frames* of the training database to converge. A single iteration on all the *frames* is required for the GMM-SVM approach to generate the UBM and the utterance GMMs. Although our MMIE training approach is very fast, because Gaussian selection is performed on the UBM, and kept fixed for all the iterations, it takes ~60 hours to produce its models starting from the standard models. The GMM-SVM approach, on the other hand, required only 2 hours to complete its job on the same dataset. Discriminative training inside the SVM training procedure is extremely fast, compared to MMI training, because it uses the *models* of the utterances, rather than their training frames. The cost of moving the language GMM supervector along its discriminative direction is negligible.

Preliminary experiments with a gender independent 2048 Gaussian GMM-SVM on the 30s tests, and with gender dependent Discriminative GMMs on the shorter duration tests, achieve performance comparable to the best ones reported in [3-4] for acoustic only systems. To enable the comparison with previous reported results, Table 2 shows the EERs obtained using pooled scores, as was usual practice before Odyssey 2006, and in parentheses the average of the EERs of each language.

6. Conclusions

A very fast, yet effective, discriminative training approach for language GMMs has been presented that exploits the information given by the separating hyperplanes estimated by a GMM-SVM classifier. Excellent results have been achieved by combining an inter-speaker variation compensation technique, the discrimination capability of the Support Vector Machines, and the accuracy of discriminative GMMs. Future work will be devoted to improving both our standard gender-dependent models and their discriminative directions.

7. References

[1] D. A. Reynolds, T. F. Quatieri and R. B. Dunn, “Speaker Verification Using Adapted Gaussian Mixture Models,” Digital Signal Processing, Vol. 10, pp. 19-41, 2000.

[2] W.M. Campbell, E. Singer, P.A. Torres-Carrasquillo, and D.A. Reynolds, “Language Recognition with Support Vector Machines,” Proc. Odyssey: The Speaker and Language Recognition Workshop, ISCA, pp. 41-44, 2004.

[3] L. Burget, P. Matejka, and J. Cernocky, “Discriminative Training Techniques for Acoustic Language Identification,” in Proc. ICASSP 2006, Vol. I, pp. 209-212, 2006.

[4] P. Matejka, L. Burget, P. Schwarz, and J. Cernocky, “Brno University of Technology System for NIST 2005 Language Recognition Evaluation,” in Proc. IEEE Odyssey 2006, San Juan, Puerto Rico, June 2006.

[5] National Institute of Standards and Technology, “NIST Speech Group Website,” http://www.nist.gov/speech/tests/lang/index.htm, 2005.

[6] W.M. 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. ICASSP 2006, Vol. I, pp. 97-100, 2006.

[7] P.A. Torres-Carrasquillo, E. Singer, M. A. Kohler, R. Greene, D. A. Reynolds, and J. R. Deller Jr., “Approaches to Language Identification using Gaussian Mixture Models and Shifted Delta Cepstral Features,” in Proc. ICSLP 2002, pp. 90-93, 2002.

[8] F. Castaldo, D. Colibro, E. Dalmasso, P. Laface and C. Vair, “Language Identification using Acoustic Models and Speaker Compensated Cepstral-Time Matrices”, to appear in Proc. ICASSP 2007.

[9] C. Vair, D. Colibro, F. Castaldo, E. Dalmasso and P. Laface, “Channel Factors Compensation in Model and Feature Domain for Speaker Recognition,” in Proc. IEEE Odyssey 2006, San Juan, Puerto Rico, June 2006.

[10] A. Canavan, and G. Zipperlen, “CALLFRIEND” available at http://www.ldc.upenn.edu.

[11] W.M. Campbell, J.R. Campbell, D.A. Reynolds, E. Singer and P.A. Torres-Carrasquillo, “Support Vector Machines for Speaker and Language Recognition”, in Computer Speech and Language, Vol. 20, pp. 210-229, 2006.

[12] M.N. Do, “Fast Approximation of Kullback-Leibler Distance for Dependence Trees and Hidden Markov Models”, IEEE Signal Processing Letters, Vol. 10, n. 4, pp. 115-118, 2003.

[13] M. Ben, “Approches Robustes pour la Vérification Automatique du locuteur par Normalisation et Adaptation Hierachique, PhD thesis, Univ. of Rennes I, 2004.