Acoustic Language Identification Using Fast Discriminative Training

Original

Availability:
This version is available at: 11583/1640722 since:

Publisher:
ISCA

Published
DOI:

Terms of use:
openAccess
This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)
Acoustic Language Identification Using Fast Discriminative Training

Fabio Castaldo\textsuperscript{1}, Daniele Colibro\textsuperscript{2}, Emanuele Dalmasso\textsuperscript{1}, Pietro Laface\textsuperscript{1}, Claudio Vair\textsuperscript{2}

\textsuperscript{1}Dipartimento di Automatica e Informatica, Politecnico di Torino, Italy
\textsuperscript{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 \cite{1,2} for acoustic LID. Discriminative training of acoustic GMMs \cite{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) \cite{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 \cite{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 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 losing 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 \cite{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 \cite{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
where new discriminative approach has been tested with gender independent models. The final score for each language includes with the set of its corresponding gender-dependent GMM best likelihood for the current utterance is selected, together sets were used for training all other types of models. The LRE-05 corpus includes seven languages and two dialects: English-American, English-Indian, 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 was split into 8172 slices of approximately 150s. The same data splits were used for training all other types of models. 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 CaliFriend [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 supervector that maps an utterance to a high dimensional space is obtained by appending the adapted mean and the diagonal covariance vectors $g$ are shared among all the GMMs, including the UBM.

Figure 1 shows an example of hyperplanes separating a class from two other classes, and their discriminative direction vectors $w_k$. 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 $\{x_i, y_i\}$, where $y_i$ is +1 and -1 for the positive and negative class targets respectively, the points $x$ that lie on the separating hyperplane satisfy the equation

$$w \cdot x + b = 0$$  \hspace{1cm} (2)

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

Figure 1 shows an example of hyperplanes separating a class from two other classes, and their discriminative direction vectors $w_k$.

4.2. GMM supervectors

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

$$f(x) = \sum_g \alpha_g \cdot N(\mu_g \cdot \sigma_g)$$  \hspace{1cm} (3)

where $N(\mu_g \cdot \sigma_g)$ is a Gaussian with mean and diagonal covariance $p$-dimensional vectors $\mu_g$ and $\sigma_g$ respectively, and $\alpha_g$ 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 $\alpha_g$ and the diagonal covariance vectors $\sigma_g$ 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]:

$$D(i, j) = \sum_g \alpha_g \sum_{g'} \left( \frac{\mu_{gp} - \mu_{gp'}}{\sigma_{gp}} \right)^2$$  \hspace{1cm} (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:

$$\bar{w}_k = \frac{w_k}{\sum_k w_k^2}$$

Figure 1: Hyperplanes separating a class from the others, and their discriminative direction vectors $w_k$. Mel frequency cepstral coefficients and their usual 7-1-3-7 Shifted Delta (SDC) features [7].
the normalized UBM supervector defines the origin of a new space, where the KL divergence is a Euclidean distance. In 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, marks the origin of this space, where the Gaussians have been estimated, each component of the standard GMM of language \( k \) will be updated according to

\[
\hat{\mu}^k_\alpha = \frac{\alpha^k \cdot \hat{\mu}'_\alpha - \mu^k_{\text{UBM}}}{\sigma^k_\alpha}
\]  

where \( \hat{\mu}'_\alpha \) is the normal to the hyperplane separating the utterance supervectors of language \( k \) from the supervectors of the other languages, and \( \alpha^k \) 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

\[
\hat{\mu}^k_\alpha \left( \alpha^k \right) = \sigma^k_\alpha \mu^k_{\text{UBM}} \left( \alpha^k \right)
\]

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. \( \hat{\mathbf{W}}^1_k \) and \( \hat{\mathbf{W}}^2_k \), in the acoustic feature space, are the rescaled components of supervector \( \mathbf{W}^k \) 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 \( \alpha^k \), 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 \( \alpha^k \).

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


