INTERSPEECH 2009 BRIGHTON

Speaker Recognition by Gaussian Information Bottleneck

Ron M Hecht[1], *Elad Noor*[2], *Naftali Tishby*[3]

[1] Department of Computer Science, Tel-Aviv University, Tel-Aviv, Israel
[2]The Weizmann Institute of Science, Rehovot, Israel
[3]School of Engineering and Computer Science, Hebrew University, Jerusalem, Israel
hadasron@gmail.com, elad.noor@gmail.com, tishby@cs.huji.ac.il

10.21437/Interspeech.2009-387

Abstract

This paper explores a novel approach for the extraction of relevant information in speaker recognition tasks. This approach uses a principled information theoretic framework - the Information Bottleneck method (IB). In our application, the method compresses the acoustic data while preserving mostly the relevant information for speaker identification. This paper focuses on a continuous version of the IB method known as the Gaussian Information Bottleneck (GIB). This version assumes that both the source and target variables are high dimensional multivariate Gaussian variables. The GIB was applied in our work to the Super Vector (SV) dimension reduction conundrum. Experiments were conducted on the male part of the NIST SRE 2005 corpora. The GIB representation was compared to other dimension reduction techniques and to a baseline system. In our experiments, the GIB outperformed the baseline system; achieving a 6.1% Equal Error Rate (EER) compared to the 15.1% EER of a baseline system.

Index Terms: Information Bottleneck method, Gaussian Information Bottleneck, Speaker Recognition, Super Vector

1. Introduction

The field of speaker recognition has developed significantly over the last few years with the introduction of the Super Vector (SV) method [1][2]. The SV approach differs from the classical Gaussian Mixture Model - Universal Background Model (UBM-GMM) [3] scheme by taking advantage of the symmetry that exists between the training and testing speech segments. Instead of using the first set only for training GMM distributions and the second set only for likelihood estimation, SVs are evaluated for both kinds of segments, and the similarity measure is given by a symmetric Euclidean distance in high-dimensional space.

SV speaker recognition systems set the UBM as the initial distribution for training GMMs. For each of the training and testing segments a GMM is estimated using maximum a posteriori (MAP) adaptation. The parameters from each GMM are then utilized for creating the SV ($\vec{x}$) for that segment, namely:

$$x_{i\cdot d+j} = \sqrt{w_i}\frac{\mu_{ij}}{\sigma_{ij}} \qquad \begin{matrix} 0\le i<g \\ 0\le j<d \end{matrix} \tag{1}$$

where w_i is the weight of the Gaussian i, μ_{ij} is the mean of the Gaussian i in the j dimension, and σ_{ij} is the standard deviation of Gaussian i in the j dimension. d is the feature space dimension and g is the number of Gaussians in a mixture. Note that the SVs have a very high dimension (equal to the products of the number of Guassians and the dimension of the features space). Given two SVs, the Euclidean distance between them is an estimation of the Kullback-Leibler (KL) divergence between their original GMM distributions [1]. In state of the art GMM-Support Vector Machine (GMM-SVM) systems [4], scores are given by the SVM, instead of using the distances (normalized using ZT-norm [5]) as done in this paper.

Unfortunately, due to lack of training data and inter-call variability, the estimation of SVs tends to be inaccurate. It can be described as if a certain amount of noise is added to the correct SV. However, some level of noise reduction can be achieved by using a procedure for reducing the dimension of the SVs [3][4][5]. In this paper, a new dimension reduction procedure is introduced, based on the GIB method, which enhances the relevant information about the speaker identity, while minimizing the overall information.

All speaker recognition systems described throughout this paper used the same acoustic features. 13 Mel-frequency cepstral coefficients (MFCC) and their first derivatives, estimated every 10ms with 25ms windows [9]. Then relevant frames were selected by an energy based Voice Activity Detector (VAD). Ultimately, a feature time warping [10] procedure was applied.

2. Methods

In this work the GIB method was applied to the speaker recognition SV models for the purpose of reducing the representation dimensionality. This section reviews the theoretical aspects of IB [11], its continuous application for multivariate Gaussian variables [12], and our new GIB solution for SV based speaker recognition.

2.1. Information Bottleneck

Each audio segment can be viewed as having two components, X and Y. X represents the acoustic characteristics of the audio signal, while Y is a representation of the classification target, in this case - the speaker identity. The amount of relevant information that X contains on Y is determined, in this method, by Shannon's mutual information between the two variables,

$$I(X;Y) = \iint_{x,y} f(x,y)\log\left(\frac{f(x,y)}{f(x)f(y)}\right)dxdy \tag{2}$$

The goal of the IB is to find a compact representation of X, denoted here by the random variable T, that on the one hand preserves as much information as possible about the speaker's identity, i.e. $\max_T I(T;Y)$, and on the other hand is as simple as possible $\min_T I(X;T)$. This procedure generalizes the classical notion of *minimal sufficient statistics*

for general random variables X and Y. The tradeoff between these two complementary criteria is obtained by introducing a positive Lagrange multiplier β and optimizing the following Lagrangian with respect to a stochastic map from X to T,

$$\arg\min_{p(t|x)}\left\{I(X;T)-\beta I(T;Y)\right\} \quad (3)$$

This optimization is similar to the Rate-Distortion tradeoff equation in lossy compression [13]. However, in the IB case an emerged distortion measure results from the information sufficiency criteria.

2.2. Gaussian Information Bottleneck

In the special case where X and Y are jointly multivariate Gaussian distributions, the IB optimization problem is turned into a generalized eigenvalue problem, as developed in [12]. The description of GIB starts with a few basic definitions. For two Gaussian random variables, X and Y, let Σ_x and Σ_y denote the covariance matrices respectively. The cross covariance matrices of X and Y is denoted by Σ_{xy} and Σ_{yx} and the conditional covariance, or canonical correlation matrix by $\Sigma_{x|y}$, defined as follows:

$$\Sigma_{x|y}=\Sigma_x-\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx} \quad (4)$$

Solving the IB for general continuous distributions can be complicated and involves an iterative solution. However, for the Gaussian case the information functions involve only these matrices and an elegant analytical solution exists. In this case, the optimal representation T is a linear projection of X on a subspace of eigenvectors of the conditional covariance matrix $\Sigma_{x|y}\Sigma_x^{-1}$. The selected eigenvectors and their relative weights are determined by the tradeoff parameter β:

$$A=\begin{Bmatrix}\left[0^T;...;0^T\right] & 0\le\beta\le\beta_1^c\\ \left[\alpha_1 v_1^T;0^T;...;0^T\right] & \beta_1^c\le\beta\le\beta_2^c\\ \left[\alpha_1 v_1^T;\alpha_2 v_2^T;0^T;...;0^T\right] & \beta_2^c\le\beta\le\beta_3^c\\ \vdots & \vdots\end{Bmatrix}, \quad (5)$$

where $\boldsymbol{A}$ is the projection matrix from X to T, $T=\boldsymbol{A}X$. The left eigenvectors are sorted according to their eigenvalues in an ascending order. The scalars α's and the critical β's, where the number of dimensions of the projection increases, are determined by the eigenvalues and eigenvectors as follows:

$$\begin{aligned}\beta_i^c&=1/(1-\lambda_i)\\ \alpha_i&=\sqrt{\frac{\beta(1-\lambda_i)-1}{\lambda_i r_i}}\\ r_i&=v_i^T\Sigma_x v_i\end{aligned} \quad (6)$$

2.3. GIB for speaker recognition

The optimal projection according to the GIB algorithm consists of the eigenvectors of the matrix $\Sigma_{x|y}\Sigma_x^{-1}$. In the speaker recognition task, the X vectors are the GMM supervectors of the voice segments (a 128*26 dimensional vector) and the Y vectors are the GMM supervectors of the speakers. The GMM supervector of a speaker was set to be the average vector of all the acoustic segments made by that speaker.

$$y_s=\frac{1}{n_s}\sum_{i=1}^{n_s}x_{si} \quad (7)$$

y_s the GMM supervector of the s^{th} speaker,

x_{si} the GMM supervector of the i^{th} segment of the s^{th} speaker,

n_s the number of segments attributed to the s^{th} speaker.

It can be seen in the following equation that under these special conditions the cross covariance matrices Σ_{xy},Σ_{yx} are both equal to the speaker covariance matrix Σ_y.

$$\begin{aligned}\Sigma_{xy}=\Sigma_{yx}&=\frac{1}{SP}\sum_{s=1}^{SP}\left(\frac{1}{n_s}\sum_{i=1}^{n_s}y_s x_{si}\right)=\\ &=\frac{1}{SP}\sum_{s=1}^{SP}\left(y_s\left(\frac{1}{n_s}\sum_{i=1}^{n_s}x_{si}\right)\right)=\frac{1}{SP}\sum_{s=1}^{SP}y_s^2=\Sigma_y\end{aligned} \quad (8)$$

where SP is the number of speakers. By using the outcome of the equation (8), the optimal projection equation can be simplified, yielding

$$\begin{aligned}\Sigma_{x|y}\Sigma_x^{-1}&=(\Sigma_x-\Sigma_{xy}\Sigma_y^{-1}\Sigma_{yx})\Sigma_x^{-1}=\\ &=(\Sigma_x-\Sigma_y\Sigma_y^{-1}\Sigma_y)\Sigma_x^{-1}=(\Sigma_x-\Sigma_y)\Sigma_x^{-1}\end{aligned} \quad (9)$$

Due to the scarcity of training data, a full covariance matrix cannot be estimated properly. To overcome this problem and simplify the estimation, in the current experiments a block-diagonal matrix was assumed [14]. This is equivalent to assuming no correlations between the values of X originate from different cepstral coefficients and no correlation between an MFCC and its derivative. Therefore, the size of each block was the number of Gaussians in the GMM. In addition, in these experiments only 128 Gaussians were used, fewer than the 1024 or 2048 commonly used in speaker recognition systems.

3. Experiments

Several sets of speaker recognition experiments were conducted in order to demonstrate the difference in performance among the several dimension reduction techniques. Three linear dimension reduction techniques were compared: Principal Components Analysis (PCA), GIB, and Linear Discriminant Analysis (LDA). In each dataset a different technique was applied and a baseline experiment was conducted as well. All the experiments were conducted using 128 Gaussians per mixture. Given that the feature space was a 26 dimensional space, the total dimension of each SV was $26\times128=3328$. Each set was composed of several experiments differentiated by the reduced dimension of the SVs. The reduced dimension values that were used spanned from no reduction at all to a dimension that was only a small fragment of the original one.

3.1. Corpora

All the experiments were conducted on the male segments of the common task of the NIST SRE 2005 [15]. The background model was estimated from the male training segments of the common task of the NIST SRE 2004. A total of about 250 segments from about 100 different speakers were used for the background training. As was mentioned earlier, in the SV approach there was a similarity between the training and testing segments, each segment was processed to a single vector in a high dimensional space. In the same way there was a similarity between T -norm and Z -norm. All the segments (training and testing) of the male data set of the common task in NIST SRE 2004 were used as T -norm and Z -norm. There were about 700 of those segments that were made by about 100 different speakers. These segments were used in the evaluation of the three matrices $(\Sigma_x, \Sigma_y, \Sigma_{x|y})$ as well.

3.2. Results

The results were compared according to two scenarios: the point with the best performance and a point with the low dimension (26*30). Performance was evaluated using Detection Error Tradeoff (DET) curves[16] and EERs [Table 1-2].

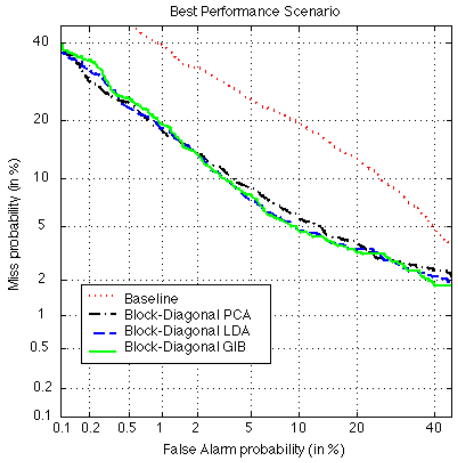

Figure 1: *Best Performance Scenario.*

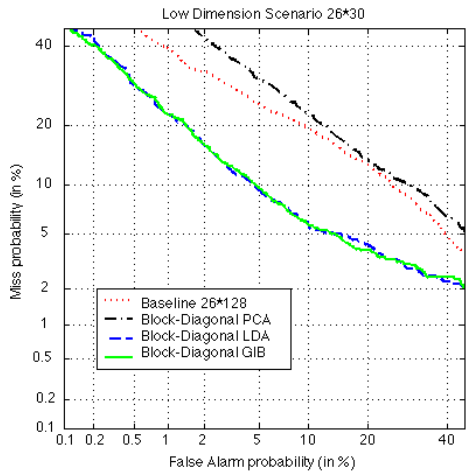

Figure 2: *Low Dimension Scenario.*

Table 1 *Best Performance Scenario.*

Algorithm	Recognition Performance (EER)	Dimension
Baseline	15.1	26*128
PCA	7.0	26*110
LDA	6.1	26*100
GIB	6.1	26*104

Table 2 *Low Dimension Scenario.*

Algorithm	Recognition Performance (EER)
Baseline	15.1
PCA	16.2
LDA	7.3
GIB	7.3

Both LDA and GIB show an improvement over the classic SV with or without PCA transformation. At their best performance point, there is no significant difference between GIB and LDA– a fact that will be explained in the discussion.

4. Discussion

According to the results, it seems that GIB is similar to LDA in most performance. We will now show that there is a theoretical equivalence between the two methods as well (section 4.1). One major advantage for GIB though, is the ability to get an explicit expression for the information residing in each of the eigenvectors, and visualizing it using a graph called the information curve (section 4.2).

4.1. Theoretic Comparison

It was decided to elaborate on the comparison of the speaker recognition version of the GIB and the LDA method, due to their convincing resemblance. Theoretical comparisons among the other techniques used, were already explored by Vogt at el. [7].

The LDA goal is to find the eigenvectors of the matrix $S_W^{-1}S_B$ that have the highest eigenvalues. S_B is the between-class scatter matrix, and is defined as Σ_y. S_W is the within-class scatter matrix and we will now prove that it is equal to $\Sigma_{x|y}$.

$$
\begin{aligned}
S_w &= \frac{1}{SP}\sum_{s=1}^{SP}\left(\frac{1}{n_s}\sum_{i=1}^{n_s}(x_{si}-y_s)(x_{si}-y_s)\right) = \\
&= \frac{1}{SP}\sum_{s=1}^{SP}\left(\left(\frac{1}{n_s}\sum_{i=1}^{n_s}(x_{si}^2)\right)-y_s^2\right) = \\
&= \frac{1}{SP}\sum_{s=1}^{SP}\left(\frac{1}{n_s}\sum_{i=1}^{n_s}x_{si}^2\right)-\frac{1}{SP}\sum_{s=1}^{SP}y_s^2 = \\
&= (\Sigma_x-\Sigma_y) = \Sigma_{x|y}
\end{aligned}
\qquad (10)
$$

Thus, the optimal projection according to the LDA algorithm is the eigenvectors of the matrix $\Sigma_{x|y}^{-1}\Sigma_y$ with the *highest* eigenvalues.

In the speaker recognition version of the GIB Y (the speaker vector) is defined as the average of all the segments of the speaker. It was shown in equation (9) that in this version of

GIB, the goal is to find the eigenvectors of the matrix $\Sigma_{x|y}\Sigma_x^{-1} = I - \Sigma_y\Sigma_x^{-1}$ that have the *smallest* eigenvalues.

Table 3.*LDA and GIB matrices.*

Algorithm	Matrix
LDA	$\Sigma_{x\|y}^{-1}\Sigma_y\vec{\nu} = \lambda\vec{\nu}$
GIB	$\vec{\omega}^T\Sigma_{x\|y}\Sigma_x^{-1} = \gamma\vec{\omega}^T$

An eigenvector and its corresponding eigenvalue of the LDA matrix are denoted by $\vec{\nu},\lambda$ and those of the GIB matrix are denoted by $\vec{\omega},\gamma$.

$$
\begin{aligned}
GIB: \quad & \vec{\omega}^T\Sigma_{x|y}\Sigma_x^{-1} = \gamma\vec{\omega}^T \\
& \Sigma_x^{-1}\Sigma_{x|y}\vec{\omega} = \gamma\vec{\omega} \\
LDA: \quad & \Sigma_{x|y}^{-1}\Sigma_y\vec{\nu} = \lambda\vec{\nu} \\
& \Sigma_{x|y}^{-1}(\Sigma_x - \Sigma_{x|y})\ \vec{\nu} = \lambda\vec{\nu} \\
& (\Sigma_{x|y}^{-1}\Sigma_x - I)\ \vec{\nu} = \lambda\vec{\nu} \\
& (\Sigma_{x|y}^{-1}\Sigma_x)\ \vec{\nu} = (1+\lambda)\vec{\nu} \\
& \frac{1}{1+\lambda}(\Sigma_{x|y}^{-1}\Sigma_x)\ \vec{\nu} = \vec{\nu}
\end{aligned}
\tag{11}
$$

By placing the eigenvector $\vec{\nu}$ of the LDA algorithm in the GIB equation, we get:

$$
\begin{aligned}
& \Sigma_x^{-1}\Sigma_{x|y}\vec{\nu} = \\
& = \Sigma_x^{-1}\Sigma_{x|y}\left(\frac{1}{1+\lambda}(\Sigma_{x|y}^{-1}\Sigma_x)\,\vec{\nu}\right) = \frac{1}{1+\lambda}\vec{\nu}
\end{aligned}
\tag{12}
$$

Therefore any eigenvector of the LDA matrix, is also an eigenvector of the GIB matrix, with some other eigenvalues. However, since there is a monotonously decreasing mapping $\lambda \to (1+\lambda)^{-1}$ between the eigenvalues, choosing the greatest for LDA, is exactly the same as choosing the smallest in GIB. Therefore, both methods will discard of exactly the same eigenvectors.

4.2. Information Curve

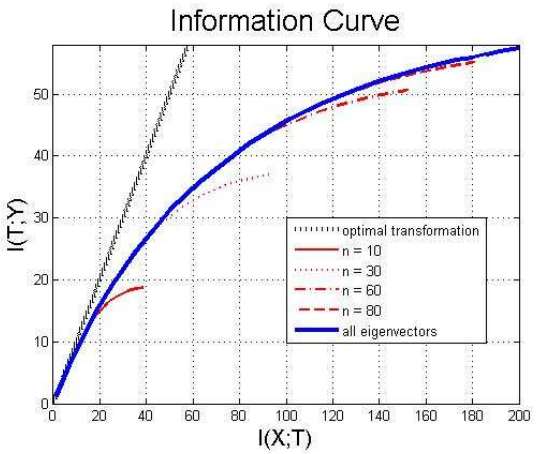

Figure 3: *Speaker Recognition Information curve. n is the number of eigenvectors per block.*

A better understanding of the recognition process can be gained by looking at the *information curve* – the dependence of the relevant information, *I(T;Y),* on the representation complexity *I(T;X)*. The easier the task - the steeper this curve should be. It can be seen that even by using a small dimension (80), the majority of the information about the speaker can be preserved. Furthermore, there is a correlation between the mutual information about the speaker and the detection performance. Still, the fact that there is no clear "knee" in the curve means that information on the speaker exists on "all scales" and all dimensions.

5. Conclusions

The GIB method in particular and the IB method in general are shown to be efficient principled methods for improving the performance of speaker recognition. The IB, through its discriminative nature, provides a subspace that contains as much information about the speaker as possible, while minimizing irrelevant acoustic information. It was additionally shown in that the classical LDA is a special case of GIB, where the relevance variable *Y* is set to be the average of the *X* vectors for each speaker. Even for this simple case, where the two variables (*X* and *Y*) are not from truly different sources, significant improvement in performance was obtained.

6. References

[1] Aronowitz, H., Burshtein, D. and Amir, A., "Speaker Indexing in Audio Archives Using Test Utterance Gaussian Mixture Modeling", in Proc. of ICSLP, 2004.

[2] Campbell, W. M., "Generalized Linear Discriminant Sequence Kernels for Speaker Recognition ", in Proc. of ICASSP, 2002.

[3] Reynolds, D. A., Quatieri, T. F. and Dunn, R. B., "Speaker Verification Using Adapted Gaussian Mixture Models", Digital Signal Processing, Vol. 10, No.1-3, 2000.

[4] Campbell, W. M., Sturim, D. E. and Reynolds, D. A., "Support Vector Machines Using GMM Supervectors for Speaker Verification", IEEE Signal Processing Letters, Vol. 13, pp. 210-229, 2006.

[5] Auckenthaler, R., Carey, M. and Lloyd-Thomas, H., "Score normalization for text-independent speaker verification systems," Digital Signal Processing, vol. 10, no. 1-3, 2000.

[6] Solomonoff, A., Quillen, C. and Campbell, W. M., "Channel Compensation for SVM Speaker Recognition", in Proc. of Odyssey, 2004.

[7] Vogt, R., Kajarekar, S. and Sridharan, S., "Discriminant NAP for SVM Speaker Recognition", in Proc. of Odyssey, 2008.

[8] Noor, E. and Aronowitz, H., "Efficient Language Identification Using Anchor Models and Support Vector Machines", in Proc. of Odyssey, 2006.

[9] "Speech processing, transmission and quality aspects (stq); distributed speech recognition; front-end feature extraction algorithm; compression algorithms," ETSI

[10] Pelecanos, J. and Sridharan, S., "Feature Warping for Robust Speaker Verification ", in Proc. of Odyssey, 2001.

[11] Tishby, N., Pereira, F., and Bialek W., ``The Information Bottleneck Method", The 37th annual Allerton Conference on Communication, Control, and Computing, 1999.

[12] Chechik, G., Globerson, A., Weiss, Y. and Tishby, N., "Information Bottleneck for Gaussian Variables", Journal of Machine Learning Research (JMLR) 6:165-188,2005.

[13] Cover, T. M. and Thomas, J. A.,Elements of Information Theory, Wiley, 2006.

[14] Aronowitz, H., Irony, D. and Burshtein, D., "Modeling Intra-Speaker Variability for Speaker Recognition", in Proc. of Interspeech, 2005.

[15] "Speaker Recognition Evaluation ", http://www.nist.gov/speech/tests/spk/.

[16] Martin A. et al., "The DET curve in assessment of detection task performance", in Proc. Eurospeech, 1997.