



A Discriminative Method for Speaker Verification Using the Difference Information

Zhenchun Lei, Yingchun Yang, and Zhaohui Wu

College of Computer Science
Zhejiang University, Hangzhou, Zhejiang, P.R.China
{leizhch, yyc, wzh}@zju.edu.cn

Abstract

In this paper, a discriminative method is proposed for speaker verification. An utterance can be mapped into a matrix by computing the difference to a codebook, and then expand the mapped matrix to a vector as the input of support vector machines for speaker verification. The Gaussian mixture model-based method is also constructed by utilizing its nature. The mapped vector indicates the utterance's fitness to the codebook. Compared with the derivative operation in the famous fisher kernel, the difference operation is used in our method. Experiments were run on the YOHO database in the text-independent case show that the new method is superior to the conventional GMM for speaker verification.

Index Terms: speaker verification, support vector machine, Gaussian Mixture Model, Vector Quantization

1. Introduction

The support vector machine (SVM) [1] is based on the principle of structural risk minimization, and has got more attention in many different fields for its superior performance. It has also been applied to speaker recognition for the discriminative training method compared with the generative models, such as Vector Quantization (VQ), Gaussian Mixture Model (GMM).

The methods using SVMs in speaker verification and speaker identification can be divided into frame-based and utterance-based. In the former, every frame is scored by the SVMs and the decision is made according to the accumulated score over the entire utterance [2]. The utterance-based approaches map an utterance into a vector as the SVM's input, and the researchers focus on how to construct a better kernel dealing with utterances having different lengths, such as fisher kernel [3] and dynamic time-alignment kernel [4].

We propose a new method which maps an utterance to a vector using the difference information to the codebook. And the GMM-based method is also constructed in the same way by utilizing the nature of the GMM. Compared with the derivative operation in the fisher kernel, the difference operation is used in our method, which reflects the fitness to the codebook. The experiments were run on the YOHO database for text-independent speaker verification.

This paper is organized in the following way: In Section 2 introduces the utterance based kernels in current years. The new discriminative method will be described in section 3. Section 4

presents the experimental results on the YOHO database for text-independent speaker verification. Finally, section 5 is devoted to the main conclusions and our future work.

2. Kernels for Utterances

SVM [4] is a binary classifier, which implements the structural risk minimization principle in statistical learning theory by generalizing optimal hyper-plane with maximum margin in two classes of data. The SVM classifier is constructed from sums of a kernel function:

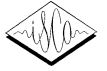
$$f(x) = \sum_{i=1}^{n_s} \alpha_i y_i K(x, x_i) + b \quad (1)$$

where n_s is the number of support vectors, $x_i, i = 1, \dots, n_s$ are the support vectors obtained from an optimization process, $y_i, i = 1, \dots, n_s$ are the corresponding targets, and α_i are the corresponding Lagrange multipliers which are found by solving an quadratic programming problem. K is the kernel function which satisfies the Mercer condition.

The kernel functions map the original input vector x into a high dimension space of features and then compute a linear separating surface in this new feature space. In practice, the use of kernel function means that an explicit transformation of the data into the feature space is not required, which is achieved by replacing the value of dot production between two data points in the input space. The kernel function defines the type of decision surface that the machines will build, and the radial basis function (RBF) kernel and the polynomial kernel are used generally.

Recently several kernels for utterances have been proposed in speaker recognition and speech applications. A well-known kernel is the fisher kernel made by Jaakkula and Haussler [3], which has been explored for speech recognition in [5] and speaker recognition in [6]. Denoting $p(x | \theta)$ is a generative model, where θ are its parameters, the mapping function is an analogous quantity to the model's sufficient statistics as following:

$$U_{\theta}(x) = \nabla_{\theta} \log(p(x | \theta)) \quad (2)$$



Each component of U is a derivative of the log-likelihood score for the input vector x with respect to a particular parameter. Campbell also propose the sequence kernel derived from generalized linear discriminates in [7, 8]. And the probabilistic distance kernel [9] is another kernel which is based on the symmetric Kullback-Leibler (KL) divergence between generative models, such as GMMs. For text depend speaker recognition, the dynamic time alignment kernel (DTAK) [4] is developed by incorporating the non-linear time alignment into the kernel function. And the pair HMM kernel [10] is similar to DTAK, but it uses a pair HMM to compute the likelihood between two utterances.

Like them, we will develop a new mapping way to deal with the variable length utterances for text-independent speaker verification.

3. New discriminative method

3.1. VQ-based method

VQ model [11] provides an effective way to describe the personal speech characters, and the decision is made depending on the scores (average distortions of whole utterance) from all models. It only considers the score, but where the accumulated score is from is ignored. So we can map an utterance into a fixed-size vector according to the score source from the codebook in the VQ model, and the accumulated differences on each codebook vector are the components of the mapped vector. On the other hand, the utterance's fitness to a codebook can be indicated in the mapped vector at the dimension-level.

The utterance, X , is denoted as a sequence of acoustic feature vectors $X = \{x_1, \dots, x_n\}$, and the vector x_i have d components. A codebook $C = \{cb_1, \dots, cb_{cn}\}$ has been gotten on one speaker's training data.

We map a frame vector x_i to a matrix according to the difference to the nearest codebook vector. For x_i , we find the vector from codebook which has the minimal distance:

$$t = \arg \min_{j=1 \dots cn} \{d(x_i, cb_j)\} \quad (3)$$

then map x_i to an matrix:

$$M(x_i) = [m_1, \dots, m_{cn}] \quad (4)$$

where

$$m_k = \begin{cases} x_i - cb_t, & k = t \\ 0, & else \end{cases} \quad (5)$$

The matrix M has $d \times cn$ size, which is same to the codebook. Now for the whole utterance X , the mapped matrix is the sum of all frames' mapped matrixes:

$$\Phi(X) = \sum_{i=1}^n M(x_i) \quad (6)$$

After expanding $\Phi(X)$ to one-dimension simply, a $d \times cn$ size vector can be got, which is the mapped result for an utterance X on the codebook C . Also we can get the linear kernel between $\Phi(X)$ and $\Phi(Y)$ directly:

$$K_{linear}(X, Y) = \sum_{i=1}^d \sum_{j=1}^{cn} \Phi(X)_{ij} \cdot \Phi(Y)_{ij} \quad (7)$$

The polynomial and RBF kernel can also be constructed in the same way.

$$K_{poly}(X, Y) = \left(\sum_{i=1}^d \sum_{j=1}^{cn} \Phi(X)_{ij} \cdot \Phi(Y)_{ij} + 1 \right)^n \quad (8)$$

$$K_{rbf}(X, Y) = \exp \left[-\frac{1}{2} \frac{\sum_{i=1}^d \sum_{j=1}^{cn} (\Phi(X)_{ij} - \Phi(Y)_{ij})^2}{\sigma^2} \right] \quad (9)$$

where n is the order of the polynomial and σ is the width of the radial basis function.

The VQ model accumulates the distortions of the whole utterance at the frame level, and our method accumulates the differences at the vector-dimension level, which is more elaborate.

Like the VQ model, the codebook is essential, which can be got by the k-means, LBG, etc. The mean vectors in the GMM can also be used to construct the codebook, and we can drive the GMM-based method.

3.2. GMM-based method

GMM [12, 13] provides a more effective way to describe the personal speech characters, and one of its powerful attributes is the capability to form smooth approximations to arbitrarily shaped densities. For a d -dimensional feature vector, x , the mixture density used for the likelihood function has the following form:

$$p(x | \lambda) = \sum_{i=1}^M w_i p_i(x) \quad (10)$$

The density is a weighted linear combination of M unimodal Gaussian densities, $p_i(x)$, each parameterized by a mean $d \times 1$ vector, μ_i , and a $d \times d$ covariance matrix, Σ_i . The parameters of GMM can be estimated using the expectation maximization (EM) algorithm.

The experiments show that the combined likelihood for x is approximate to the max component likelihood when M is small, and the top 5 components are used in the Universal Background Model (UBM) which has 1024 or 2048 components.

$$p(x | \lambda) \approx \max_i w_i p_i(x) \quad (11)$$



Denote the m is the index of component having max likelihood, the log-likelihood score can be expressed as a distance function:

$$\begin{aligned} \log(p(x|\lambda)) &\approx \log(w_m P_m(x)) \\ &= \log(w_m) - \frac{1}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_m|) - \frac{1}{2} \frac{|x - \mu_m|^2}{|\Sigma_m|} \quad (12) \\ &= A + B|x - \mu_m|^2 \end{aligned}$$

Both scores in this GMM and VQ are computed according to the distance, so we can construct the GMM-based method in the same way. Obviously, the mean vectors can be used as the codebook directly. And there are two improved sides by utilizing the nature of GMM: First, for x_i , we can select the component from all components in the GMM which has the maximal likelihood instead of the nearest Euclidean distance in the VQ-based method:

$$t = \arg \max_{j=1 \dots M} \{w_j p_j(x_i)\} \quad (13)$$

Secondly, the normalization operation is used to map x_i to a matrix $M(x_i) = [m_1, \dots, m_M]$. Naturally the covariance matrixes in the GMM parameters can be used as the normalization factors:

$$m_k = \begin{cases} \frac{x_i - \mu_t}{\text{sqrt}(\Sigma_t)}, & k = t \\ 0, & \text{else} \end{cases} \quad (14)$$

The matrix M has $d \times M$ size and it reflects the weighted distance between x_i and its nearest Gaussian component at the dimension-level.

For text independent speaker verification, the fish kernel is studied in some literatures and is combined with the GMM in general. Like it, our method is also a new way to map utterances into vectors and reflect the utterances' fitness to the codebook, but there are two merits in theory. First, the base idea in ours is more simply which only uses the difference operation instead of the derivative operation. Secondly, although the fisher kernel is correlative to the generative model, it is not feasible to combine with the VQ model because we can't get the parameters derivative. Our method is developed from the VQ model, and so it is well to combine with both the VQ and GMM.

4. Experiments

4.1. Setup

Our experiments were performed using the YOHO [14] database, which consists of 138 speaker prompted to read combination lock phrases. Every speaker has four enrollment sessions with 24 phrases per session and 10 verify sessions with 4 phrases per session. The features are derived from the waveforms using 12th order MFCC on a 20 millisecond frame every 10 milliseconds and deltas computed making up a thirty

two dimensional feature vector. Mean removal, preemphasis and a hamming window were applied. And energy-based end pointing eliminated non-speech frames.

The SVM is constructed to solve the problem of binary classification. For the N-class problem, the general method is to construct N SVMs. Training SVMs rely on quadratic programming optimizers, and the SMO [15] algorithm was used in our experiments.

The whole database was divided into two parts. The first parts of speakers, labeled 101 to 174, were trained respectively on the same imposters who were labeled 175 to 277. Then the target speaker's and the imposters' utterances were been mapped according to the models, and the support vector machines were trained using these mapped vectors as the inputs.

There are some score normalization methods for speaker verification, and we use the cohort approach, which uses a set of cohort speaker who are close to the target speaker. The size of the cohort in our experiments is 1.

4.2. GMM order

The first experiment was run using the GMM with different order and GMM/SVMs with the corresponding difference information. Table 1 shows the Equal Error Rate (EER) and the figure 1 shows some DET curves

Table 1: the EER for speaker verification

| Model | EER (%) | | | |
|----------------|---------|------|------|------|
| | 8 | 16 | 32 | 64 |
| GMM | 13.7 | 9.2 | 6.0 | 4.2 |
| GMM/SVM | 3.0 | 1.8 | 1.5 | 1.3 |
| GMM+cohort | 1.4 | 0.66 | 0.36 | 0.24 |
| GMM/SVM+cohort | 0.89 | 0.40 | 0.17 | 0.09 |

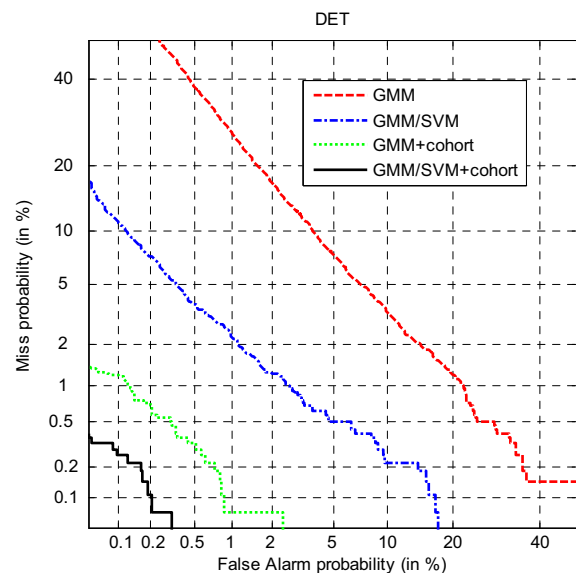
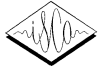


Fig.1. DET plots of the 32 order GMM approach and the corresponding GMM/SVM method



The SVMs were trained using the linear kernel function. It could be seen that the GMM/SVM gave a marked improvement compared with the same order GMM. After cohort normalization, GMM/SVM could get relative reductions of 36~63% compared with the same order GMM.

4.3. VQ-based and GMM-based methods

Table 2 shows the GMM/SVM performance compared with the VQ/SVM method.

Table 2: the EER using the VQ/SVM and the GMM/SVM for speaker verification

| Model | EER (%) | | | |
|----------------|---------|------|------|------|
| | 8 | 16 | 32 | 64 |
| VQ/SVM | 3.6 | 2.8 | 2.3 | 1.8 |
| GMM/SVM | 3.0 | 1.8 | 1.5 | 1.3 |
| VQ/SVM+cohort | 1.8 | 0.9 | 0.5 | 0.36 |
| GMM/SVM+cohort | 0.89 | 0.40 | 0.17 | 0.09 |

We can see that the performance of the GMM/SVM is better than the VQ/SVM's clearly. After the cohort normalization, the GMM/SVM can get relative reductions of 51~75% compared with the same order VQ/SVM.

4.4. Kernel function

The kernel functions decided the type of decision surfaces and the previous experiments were run using the linear kernel function. We evaluated the performances of the polynomial function and the radial basis function also.

Table 3 shows the result. According to the GMMs with 32 Gaussian components, the SVMs were trained using the appropriate parameters. Both the polynomial function and the radial basis function could improve the performances in some measure, and there was no clear difference between them.

Table 3: the EER with the different kernel functions for speaker verification

| Kernel type | EER (%) | |
|---------------------|-----------|--------|
| | No cohort | cohort |
| linear | 1.5 | 0.17 |
| polynomial (n=2) | 1.6 | 0.13 |
| rbf ($\sigma=50$) | 1.6 | 0.13 |

5. Conclusion

A new discriminative method was proposed in this paper adopting the difference information to the codebook and the GMM-based method was driven by combing the GMM's parameters for speaker verification. The base idea is mapping an utterance to a fixed-size vector thought computing the difference to a codebook. And the GMM's nature was adopted for improving the performance by finding the maximal likelihood component and the covariance normalization. The experiments on the YOHO database show that our method can be superior to the conventional GMM for speaker verification. In the future, we will research the high order deviation, and the way to combine with the GMM will also be improved by taking full advantage of its nature.

6. Acknowledgements

This work is supported by National Science Fund for Distinguished Young Scholars60525202, Program for New Century Excellent Talents in University NCET-04-0545 and Key Program of Natural Science Foundation of China 60533040.

7. References

- [1] V.Vapnik. Statistical Learning Theory. John Wiley and Sons, New York, 1998
- [2] V.Wan, W.M.Campbell, "Support Vector Machines for Speaker Verification and Identification," in Proc. Neural Networks for Signal Processing X, pp.775-784, 2000
- [3] T.S.Jakkola and D.Hausler. "Exploiting generative models in discriminative classifiers," In Advances in Neural Information Processing System 11, M.S.Kearns, S.A.Solla, and D.A.Cohn, Eds. MTT Press, (1999)
- [4] Hiroshi Shimodaira, Kenichi Noma, Mitsuru Nakai and Shigeki Sagayama, "Dynamic Time-Alignment Kernel in Support Vector Machine," NIPS, pp.921-928, 2001
- [5] Nathan Smith, Mark Gales, and Mahesan Niranjan, "Data-dependent kernel in SVM classification of speech patterns," Tech.Rep. CUED/F-INFENG/TR.387, Cambridge University Engineering Department, 2001
- [6] Shai Fine, Jiri Navratil, and Ramesh A.Gopinath, "A hybrid GMM/SVM approach to speaker recognition," in Proc. of the International Conference on Acoustics, Speech, and Signal Processing, 2001
- [7] W.M.Campbell, "Generalized Linear Discriminant Sequence Kernel for Speaker Recognition," in Proc. of the International Conference on Acoustics, Speech, and Signal Processing, pp.161-164, 2002
- [8] W.M.Campbell, "A SVM/HMM System for Speaker Recognition," in Proc. of the International Conference on Acoustics, Speech, and Signal Processing, 2003
- [9] P.J.moreno and P.P.Ho, "A New SVM Approach to Speaker Identification and Verification Using Probabilistic Distance Kernels", in Eurospeech, 2003
- [10] Vincent Wan, Steve Renals, "Evaluation of Kernel Methods for Speaker and Identification," in.Proc. of the International Conference on Acoustics, Speech, and Signal Processing, 2002
- [11] F. K. Soong et al., "A vector quantization approach to speaker recognition," in.Proc. of the International Conference on Acoustics, Speech, and Signal Processing,, 1985
- [12] D.A.Reynolds and R.C.Rose, "Robust text-independent speaker identification using Gaussian mixture speaker models," IEEE Trans. Speech Audio Processing, vol.3, pp.72-83, 1995
- [13] D.A.Reynolds, T.Quatieri, and R.Dunn, "Speaker verification using adapted Gaussian mixture models," Digital Signal Processing, vol.10, no.1-3, 2000
- [14] J.P.Campbell Jr., "Testing with the YOHO CD-ROM voice verification corpus," in.Proc. of the International Conference on Acoustics, Speech, and Signal Processing, 1995
- [15] J.Platt, "Fast training of SVMs using sequential minimal optimization," Advances in Kernel Methods: Support Vector Learning, MIT press, Cambridge, MA, 1999