Combination of Pitch and MFCC GMM Supervectors for Speaker Verification

Wei Huang[1], Jianshu Chao[2], Yaxin Zhang[1]
[1] *Motorola China Research Center, Shanghai, China*
[2] *Shanghai Jiao Tong university, Shanghai, China*
[1] *{A5901C, A12586}@motorola.com* [2] *charles_js@sjtu.edu.cn*

Abstract

A large majority of speaker verification systems are based on frame-level acoustic features, such as Mel Frequency Cepstral Coefficients (MFCCs) which characterize the vocal tract contribution. The most commonly used statistical GMM-UBM classifier models the distribution of MFCCs quite well. Pitch is one of the most important features which characterize speaker-dependent vocal fold vibration rate. It can complement the vocal tract information as source information. Although the source information is supposed to follow a lognormal distribution, the discriminative Support Vector Machine (SVM) is more suitable for pitch classification. In this paper, firstly we exploit GMM-UBM and SVM to the frame-level pitch vectors. Then we put the state-of-the-art GMM supervectors concept to the pitch feature vectors and experiment shows a promising result. And the combination of two feature type GMM supervectors systems gains much better performance. All experiment results are obtained on the NIST 2001 Speaker Database.

1. Introduction

Mel Frequency Cepstral Coefficients (MFCCs) are one of the most popular features used in the UBM-GMM system. MFCC coefficients are supposed to characterize the vocal tract contribution accurately. The Gaussian Mixture Model is a statistical classifier which represents the distribution of speech feature vectors. MFCC coefficients fit well with the idea of Gaussian distribution. Nevertheless, they are easily affected by channel variability and fail to capture the longer-range information and source information that also resides in the signal.

The pitch is used as prosodic features to characterize the source contribution (glottis). The distributions of frame-level pitch values have been used in a number of studies [1]. Statistics of pitch are used as prosodic features in speaker recognition system and have proven to be more robust than cepstral features to channel mismatches [2]. In [3], the author showed that the lognormal distribution fits pitch histograms much better than the normal distribution. Arcienega and Drygajlo [4] have presented a statistical approach using pitch dependent GMMs.

However GMMs have the disadvantage of discrimination. SVMs are one of the best discriminating models among Machine learning and they have been successfully exploited to a number of applications such as handwriting, face identification, bioinformatics. The state-of-the-art approach is to use a GMM supervector consisting of the stacked means of the mixture components [5]. Linear classification techniques are applied in the potentially high-dimensional space. From this concept, we compute frame-level MFCCs and pitch values and then turn these frame-level vectors into conversation-level supervectors. At first, we examine these systems separately. Since pitch can complement the vocal tract information, we make a combination of the two feature type systems on the score-level.

2. Baseline GMM-UBM system

Since NIST 1996 Speaker Recognition Evaluations (SRE), the Gaussian Mixture Model-Universal Background Model (GMM-UBM) speaker verification system has become dominating system because of its excellent performance in text-independent speaker recognition tasks. A GMM which used in speaker recognition applications represents multivariate probabilistic densities of speech feature vectors. The probabilistic model makes it suitable for unconstrained text-independent applications.

The UBM is generally a GMM trained from a quite large pool of speech database to represent the speaker-

independent distribution of features including various speakers, category of language, handset types, ambient environment, channel variability, and so on. In the GMM-UBM system, we adapt the parameters of the UBM using the speaker's enrollment speech and Maximum A Posteriori (MAP) estimation to derive the corresponding speaker model. During testing an unknown utterance, the system calculates the likelihood ratio of producing the unknown utterance between the enrollment model and UBM. The log-likelihood ratio formula is $\Lambda(x) = \log p(X \mid \lambda_{hyp}) - \log p(X \mid \lambda_{ubm})$. Figure 1. shows the framework of likelihood ratio-based speaker verification system [6].

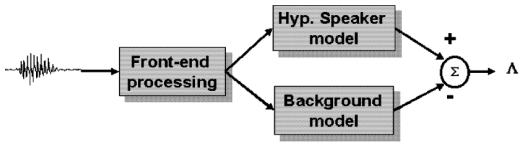

Figure 1. Likelihood ratio-based speaker verification system

3. Support vector machines

An SVM is a two-classifier based on a hyperplane boundary. The decision boundary should be far away from the data of both classes as possible. The basic fundamental of SVM is to project input vectors into a possibly infinite-dimensional space in which a hyperplane can separate each classes linearly. Given speaker feature vectors $x_i \in X$ and let $y_i \in \{1, -1\}$ be the class label of x_i. SVM uses a mapping function $\Phi(x)$ to input feature space X and an SVM discriminative function is given by

$$f(x) = \sum_{k=1}^{M} \lambda_k y_k < \phi(x), \phi(x_k) > + b = \sum_{k=1}^{M} \lambda_k y_k K(x, x_k) + b \quad (1)$$

Here, f(x) is under these constraints $\sum_{k=1}^{M} \lambda_k y_k = 0$, $\lambda_k > 0$. $K(x_i, x_j)$ is the kernel function that constrained to satisfy the Mercer condition [7]. y_i are the ideal outputs, either +1 or -1, depending on whether the corresponding support vector is in class +1 or -1. M is the number of support vectors and λ_k are trained through a software tool libSVM [8]. For testing an unknown speaker utterance, the class decision is decided by whether the score f(x) is above or below a threshold using libSVM.

The lognormal distribution fits pitch histograms much better than the normal distribution [3]. Nevertheless, the variability of pitch changes a little because the pitch is speech fundamental frequency which reflects vocal fold vibration. It's a speaker-dependent physical property determined by the size, mass and stiffness of the speaker's vocal folds. Figure 2. shows the distributions of pitch and MFCCs in the two-dimensional space. The statistical property of Pitch+Δ Pitch (left) is less obvious than the MFCCs' (right). SVM is a discriminating model based on the principle of structural risk minimization. It doesn't concern the statistical distribution of one speaker feature vectors. Using the mapping of low-dimensional space into a high-dimensional space, SVM shows its strength in classification problem such as the pitch classification.

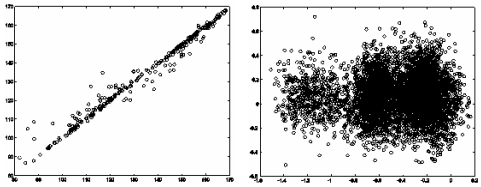

Figure 2. One speaker's pitch and MFCCs distribution in two-dimensional space

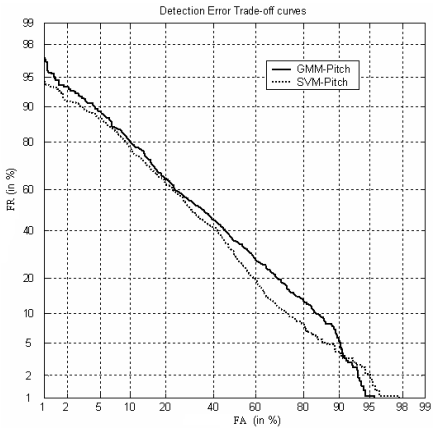

Figure 3. Comparison between GMM and SVM systems using pitch features only

We followed the method in [9] using GMM to present the distribution of pitch features. Fig. 3 show the comparison between GMM and SVM using pitch features only. This result is what we have done before [10]. We can see that the SVM yields better classification result than GMM system. But the EER of the two systems are about 30% high. They are both worse than the MFCCs based system. The pitch features contain less discriminative information for every speaker and they can't be used as speaker verification system lonely.

4. GMM supervectors system

The general probability density function (pdf) of an N-component Gaussian mixture model universal background model (GMM-UBM) is defined as

$$p(x) = \sum_{i=1}^{N} w_i \mathrm{N}(x; m_i, \Sigma_i) \quad (2)$$

where N( ) is the Gaussian density function and $\{w_i, m_i, \Sigma_i\}$ is the set of parameters in the model. They respectively represent the weights (constrained to $\sum_{i=1}^{N} w_i = 1$), the d-dimensional mean vectors and the d $\times$ d covariance matrices.

Given a speaker utterance, the GMM-UBM is adapted by Maximum A Posteriori (MAP) adaptation [6] to produce the corresponding speaker model. Generally, only the means m_i of Gaussian components are adapted. From this adapted model, we obtain a GMM supervector by concatenating all GMM Gaussian mean vectors. GMM supervector can be thought of as a mapping between an utterance and a high-dimensional vector. The process is shown in Figure 4.

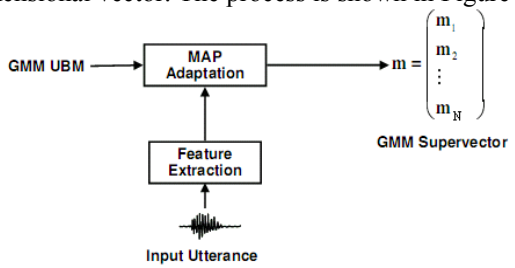

Figure. 4 GMM Supervectors concept [5]

The GMM supervector SVM is based on the adapted model means. Instead of using the MFCC features or pitch value features directly, it uses the adapted Gaussian means which form a supervector as features. The supervector is then treated as an SVM classifier input, similar to the normal SVM classification.

The main design component is an SVM kernel, which is an inner product in the supervector space. Since inner products induce distance metrics and vice versa, the basic goal in SVM design is to find an appropriate metric in the SVM feature space relevant to the classification problem. In [5], the author proposed a novel linear kernel.

$$\begin{aligned} K(utt_a, utt_b) &= \sum_{i=1}^{N} w_i u_i^a \sum_i^{-1} (u_i^b)^t \\ &= \sum_{i=1}^{N} (\sqrt{w_i} \sum_i^{-\frac{1}{2}} u_i^a)(\sqrt{w_i} \sum_i^{-\frac{1}{2}} u_i^b)^t \end{aligned} \quad (3)$$

The kernel in (3) satisfies the Mercer condition and it is linear in the GMM supervector space since it concatenates all GMM mean vectors and normalized by corresponding standard deviations.

5. Fusion of pitch and MFCC GMM supervectors systems

Considering the property of pitch, SVM classifier will be much suitable. And in the former GMM system, although the frame-level pitch features convey fundamental information, such statistics don't capture dynamic information about pitch contours and thus not viewed as high-level. For these reasons, we exploit the same GMM supervectors method to pitch features. These supervectors of pitch features can be thought of a mapping between an utterance and conversational level speaker fundamental source information.

The GMM supervectors system using MFCC features has proven to be an effective system. In our experiments, the traditional GMM system, SVM system and supervectors SVM system using simple pitch features can't perform well in speaker verification tasks. But the fusion of the two feature type systems will combine the advantage of them. Therefore, we treat MFCCs based system as the main system for its greater performance and accurate characterization of vocal tract configuration and pitch features based system as reference system. Our fused GMM supervectors system is combined on the score-level. Fig. 5 shows our fused system frame.

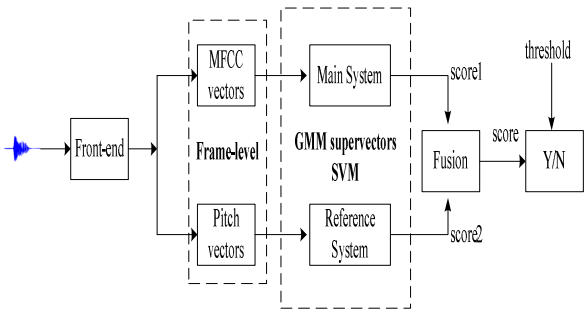

Figure 5. Fused system frame

6. Experiments

6.1. Database

All experiments are performed on the NIST 2001 SRE database in the task of one-speaker detection [11]. The development set contains 38 male and 22 female speakers and the evaluation set contains 74 male and 100 females. For training a model, there is approximately 2 minutes enrollment speech for each speaker. For testing a system, there are 850 male and 1188 female verification utterances lasting about 60 seconds. And there are 1 target and 10 imposter trials of the same sex for each enrolled speaker. Performance is evaluated based on all target scores and imposter scores. We use equal error rate (EER) and minimum Detection Cost Function (DCF) as metrics for performance evaluation.

6.2. Front-end processing

12-dimensional Mel-frequency cepstral coefficients (MFCCs) are extracted every 10ms using a Hamming window of 20ms. First-order deltas and second-order deltas are appended to the cepstral vectors forming a 36 dimensional feature vectors. Then Cepstral mean subtraction (CMS) is applied to the MFCCs to remove linear channel effects.

We extract log pitch and log energy values every 25ms, and a shift of 10ms. The pitch is based on the computation of autocorrelation function estimated from cochlear filter banks. Here, we consider the log energy distribution to satisfy normal distribution. Then the first-order deltas and second-order deltas are also appended to the two values forming a 6 dimensional feature vectors. We don't use average pitch or pitch contours, so our pitch feature vectors are based on frame-level.

6.3. GMM-UBM baseline systems

In our experiments, the MFCCs and pitch GMMs are modeled respectively by 512 and 64 mixtures of Gaussians and diagonal matrices are used. For each of the two type features, two gender-dependent UBMs (male and female background models) are trained from NIST 2001 development database respectively. In another word, there are totally four UBM needed to be trained. Then each target speaker model is created by only adapting the matching gender UBM means using Maximum A Posteriori (MAP) adaptation.

6.4. Fused GMM supervectors system

The system produces target GMM supervector as +1 class, and other 173 imposter model supervectors as -1 class when training an SVM model. The processes of producing main system and reference system are almost the same. When testing an unknown utterance, we use the testing feature vectors to produce GMM supervector as input of SVM. Then we combine the main system and reference system on the score level. The weights of the two systems are selected through experimental method. This is not the most optimal selection, but it does work.

7. Results and conclusions

Firstly, we compare the systems of baseline GMM and supervectors SVM only using pitch features. The latter system gains 26.31% higher profit than the former one in EER. The result shows the idea of projecting the frame-level pitch to a conversational level using supervectors SVM method works well. The pitch supervectors represent the fundamental source better than the frame-level pitch vectors. The author in [1] takes a broad view and includes as higher-level any features that involve either linguistic information or information at longer time spans than used in frame-based system. In their experiment, their most successful longer-range system is based on the conditioned syllable-based prosody sequence. Here in our experiment, pitch supevectors are derived from the frame-level at shorter time we may not treat them as high-level. Our future work should compare the difference of pitch supervectors system and high-level such as pitch contour systems which capture longer-range information.

Table 1. Comparison between GMM and supervectors SVM systems using pitch vectors

System	EER	DCF ($\times 10^{-3}$)
GMM-UBM(pitch)	29.04%	99.4
GMM supervectors SVM(pitch)	21.4%	85.0

Starting from the baseline GMM-UBM system, we examine the supervectors SVM system and our pitch and MFCCs fused system. Obviously, the fused system yields the best result. The frame-level derived pitch supervectors and MFCC supervectors complement each other in the supervector space since the source and vocal tract are taken into account. From the table and EER curves, we can see that this fusion approach is promising. And our weights of two systems are selected by experimental try. The selection can be optimized by using a full search method.

The pitch presents the fundamental source and is more resistant to channel variation. Meanwhile the MFCCs are easily affected in acoustic conditions. If we remove the unwanted variation in the main system by using compensation methods such as nuisance

attribute projection [12], the supervector space of the main system could be more suitable in the whole fused system. Our future work will focus on that.

Table 2. Comparison between GMM, supervectors SVM and fused systems.

System	EER	DCF ($\times 10^{-3}$)
GMM-UBM(MFCCs)	8.68%	37.6
GMM supervectors SVM(MFCCs)	7.21%	30.0
Fused SVM system (pitch and MFCC supervectors)	6.67%	27.4

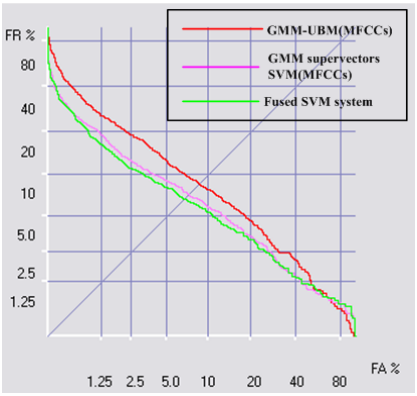

Figure 6. EERs of three systems

8. References

[1] lizabeth Shriberg, "Higher-Level Features in Speaker Recognition", *Lecture Notes in Computer Science*, Volume 4343, 2007

[2] J. Carey, E.S. Parris, H. Lloyd-Thomas, and S.Bennett, "Robust prosodic features for speaker identification". *In NIST Speaker Recognition Workshop*, March 1996

[3] M. Kemal SSnmez, Larry Heck, Mitchel Weintraub, and Elizabeth Shriberg. "A lognormal tied mixture model of pitch for prosody-based speaker recognition", *In Proceedings from Eurospeech*, volume 3, pages 1391-1394, 1997.

[4] Arcienega M. and Drygajlo A, "Pitch-dependent GMMs For Text-Independent Speaker Recognition Systems", *In Eurospeech'01*, 2001, Scandinavia.

[5] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, "Support Vector machines using GMM supervectors for speaker verication," *IEEE Signal Processing Letters,* 2005.

[6] D. Reynolds, T. Quatieri, and R. Dunn, "Speaker Verification using Adaped Gaussian Mixture Models," *Digital Signal Processing,* vol. 10, no. 1-3, pp. 19-41, 2000.

[7] Nello Cristianini and John Shawe-Taylor, Support Vector Machines, Cambridge University Press, Cambridge, 2000.

[8] Chih-Chung Chang, Chih-Jen Lin. "LIBSVM: a Library for Support Vector Machines," "http://www.csie.ntu. edu.tw/~cjlin/libsvm/index.html".

[9] Hassan Ezzaidi and Jean Rouat "Pitch and MFCC dependent GMM models for speaker identification systems", *Electrical and Computer Engineering, Canadian Conference* on Volume 1, Issue , 2-5 May 2004 Page(s): 43 - 46 Vol.1

[10] Wei Huang, "Text-independent speaker recognition based on GMM/SVM and multi-system fusion", Ph.D. dissertation (written in Chinese), USTC, China, 2004.

[11] The NIST year 2001 speaker recognition evaluation plan. http://www.nist.gov/speech/tests/spk/2001/doc.

[12] W. Campbell, D. Sturim, D. Reynolds, and A. Solomonoff, "SVM Based Speaker Verification using a GMM Supervector Kernel and NAP Variability Compensation," *in ICASSP*, vol. 1, 2006, pp. 97-100.