Front. Electr. Electron. Eng. China 2010, 5(1): 72–76
DOI 10.1007/s11460-009-0073-3

RESEARCH ARTICLE

Rongyan WANG, Gang LIU, Jun GUO, Yu FANG

Multi-class classifier of non-speech audio based on Fisher kernel

Abstract Traditional multi-class classification methods based on Fisher kernel combine generative models such as Gaussian mixture models (GMMs) of all the classes together. However, the combination generates high dimensional feature vectors and leads to large computation. In this paper, a new classification method is proposed. This method adopts an intelligent feature space selection strategy by clustering similar Gaussian mixtures in order to reduce the feature dimensions. Audio classification experiments show that the proposed method is more accurate and effective with less computation compared with traditional methods.

Keywords Fisher kernel, support vector machine (SVM), Gaussian mixture model (GMM), mixture clustering

1 Introduction

Researchers of automatic speech recognition have addressed the problem of interpreting speech by machine for decades. Their efforts have led to a considerable understanding of the domain and numerous practical applications. However, researches on content-based classification and retrieval of non-speech audio files are relatively new.

In recent years, although many researches have appeared on audio classification and retrieval, e.g., Refs. [1] (http://www.musclefish.com) and [2] (http://www.comparisonics.com), they tend to focus on matching test sounds into a limited number of predefined categories, such as music, applause, speech, etc. In Refs. [3–5], audio files are classified into three types: speech, music and others. There are also some researches about multi-classification of non-speech audio files [6–9]. In Ref. [6], Gaussian mixture model (GMM) is used to classify and retrieve the audio files including 11 broad kinds. At the same time, current researches tend to use support vector machine (SVM) because SVM often outperforms generative models in classification. SVM demands features with fixed length. In Refs. [7–9], the mean and the standard variance of the features extracted from each frame of an audio clip with 1 s at least are used for audio analysis. Despite the good results obtained, an unavoidable loss of information exists when frame-level features are transformed into statistical clip-level features. There is another problem that the results tend to be worse when the clip is shorter than 1 s.

As a result, a special kernel called Fisher kernel, which is based on generative model, has been applied to generate fixed-length features. It is already used in bio-sequence analysis, speech recognition, and speaker identification [10–13]. Unlike previous schemes [14,15], the proposed scheme trains GMM for each class first, and all the models are merged to a single one. Then, Gaussian mixture clustering is used to generate new effective and fixed-length features in Fisher score space derived from this new model. Finally, SVM is used to classify the audio files into six classes. The proposed method is described in Sect. 2 and Sect. 3 in detail.

The organization of this paper is as follows: In Sect. 2, we introduce the Fisher kernel methodology and strategies for multi-class classification. The feature reduction strategy of Gaussian mixture clustering is discussed in Sect. 3. The results of the experiments are reported in Sect. 4. Section 5 gives the conclusion.

Received July 11, 2009; accepted September 6, 2009

Rongyan WANG (✉), Gang LIU, Jun GUO, Yu FANG
Pattern Recognition and Intelligent System Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China
E-mail: wangrongyan8232@163.com

2 Theory of Fisher kernel and strategies for multi-class classification

Generative probability models such as hidden Markov model (HMM) or GMM provide principled ways of treating missing information and dealing with variable

length sequences. Discriminative methods such as SVM enable us to construct flexible decision boundaries and often result in a classification performance superior to that of model-based approaches. However, discriminative methods demand that each audio file has the same length, while the audio files we can get are always different in length.

2.1 GMM

GMM is a widely used supervised classification technique that improves over a single Gaussian distribution to accommodate a broader, more complex range of distributions using a combination of simple components. They can effectively model distributions where the data points originate from separate clusters but membership is unknown.

To model an acoustic feature vector $\boldsymbol{X}$ (of dimension D), the distribution is described by

$$p(\boldsymbol{X}) = \sum_{i=1}^{K} \pi_i p(\boldsymbol{X}|\boldsymbol{\theta}_i), \tag{1}$$

where K is the number of mixture components, π_i is the probability that component i contributes to modeling the data, and $\boldsymbol{\theta}_i$ is the parameter of component i. In the case of GMMs, $\boldsymbol{\theta}_i = \{\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i\}$ and the probability density function (PDF) is a multivariate Gaussian distribution:

$$p(\boldsymbol{X}|\boldsymbol{\theta}_i) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}_i|^{1/2}} \exp\left[-\frac{1}{2}(\boldsymbol{X}-\boldsymbol{\mu}_i)^{\mathrm{T}}\boldsymbol{\Sigma}_i^{-1}(\boldsymbol{X}-\boldsymbol{\mu}_i)\right], \tag{2}$$

where $\boldsymbol{\mu}_i$ is D-dimensional mean vector and $\boldsymbol{\Sigma}_i$ is $D\times D$ covariance matrix of component i. Hence, for each training case, there are two sets of parameters to estimate, the mixing weights π_i and the parameters $\boldsymbol{\theta}_i$. An appropriate value for K must also be determined.

2.2 Fisher kernel

Fisher kernel is proposed as a methodology to map variable length sequences to a new fixed dimension feature vector space [11]. This new feature space is called the Fisher score space. On this score space, any discriminative classifier can be used to perform a discriminative training. The main idea of Fisher kernel is to combine generative models with discriminative classifiers to obtain a robust classifier that has the strength of each approach. Since each Fisher score space is based on a single generative model, the new feature space is assumed to be suitable for classification problems in nature. The features of fixed length are obtained as follows.

Given a set of labeled audio feature sequences $\boldsymbol{X} = \{\boldsymbol{X}_i | i = 1,2,...,N\}$, each composed of a different number of vectors. Consider a class of static model $\boldsymbol{\theta}$, $\boldsymbol{\theta} = \{\boldsymbol{\theta}_i | i = 1,2,...,M\}$, for elaboration. $p(\boldsymbol{X}_i|\boldsymbol{\theta})$ evaluates the probability of $\boldsymbol{X}_i$ measured by model $\boldsymbol{\theta}$. Fisher score vector is the gradients of the logarithm of $p(\boldsymbol{X}_i|\boldsymbol{\theta})$ with regard to each parameter $\boldsymbol{\theta}_i$, that is,

$$\begin{aligned} U_{\boldsymbol{X}_i} &= \nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{X}_i|\boldsymbol{\theta}) \\ &= \left[\frac{\partial \ln p(\boldsymbol{X}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1}, \frac{\partial \ln p(\boldsymbol{X}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_2}, ..., \frac{\partial \ln p(\boldsymbol{X}_i|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_M}\right]^{\mathrm{T}}. \end{aligned} \tag{3}$$

In this paper, the generative model used is GMM, so $\boldsymbol{\theta}_i$ can be weight $p(l)$, each component of the mean vector $\boldsymbol{\mu}_l$ and the diagonal covariance matrix $\boldsymbol{\Sigma}$. For example, if $\boldsymbol{\theta}_i = \boldsymbol{\mu}_l$, the Fisher score vector is

$$\nabla_{\boldsymbol{\mu}_l} \ln p(\boldsymbol{X}_i|\boldsymbol{\theta}) = \sum_{j=1}^{n_i} \frac{p(l)p(\boldsymbol{X}_{ij}|\boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)}{p(\boldsymbol{X}_{ij}|\boldsymbol{\theta})} \boldsymbol{\Sigma}_l^{-1}(\boldsymbol{X}_i - \boldsymbol{\mu}_l), \tag{4}$$

where j is the frame number of the sample $\boldsymbol{X}_i$.

Obviously, the dimension of Fisher score vector is decided by the mean vector $\boldsymbol{\mu}_l$ and the number of mixtures. In fact, Fisher score can be interpreted in the following way. When a generative model is trained by maximum likelihood (ML) criterion, it uses the same set of derivatives to compute how close it is to the local extreme. Another motivation of using Fisher score is that the gradient of the log-likelihood can capture the generative process of the whole sequence better than just a posterior probability. Furthermore, it was shown from Eq. (4) that, under the condition that the class variable is a latent variable in the probability model, the learning machines, using Fisher kernel are asymptotically at least as good as making decisions based on the generative model itself (maximum a posteriori). Therefore, we can use the new features in Fisher score space to train discriminative classification and get a good result.

2.3 Multi-class classification strategy

In the frameworks proposed in Ref. [11], Fisher scores are computed from the log-likelihood of a single statistical model representing one class or both competing classes. However, when Fisher scores are applied to multi-class classification problems, the feature vectors of Fisher scores based on one model may lose some discriminative information. The block diagram of our approach to solve the problem is shown in Fig. 1.

In Fig. 1, after GMMs of each class are trained, all the models are combined to a particular statistical model. The new model is obtained as follows: re-weigh all the weights $p(l)$ of each class to make sure they add up to 1.0 and all sets of Gaussian mixtures are combined to generate the mixtures of the new model. The number of mixtures in the new signal model is the sum of the number of mixtures in each class. Then, use mixture clustering method to merge

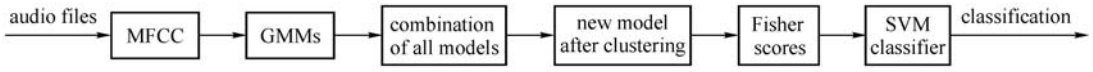

Fig. 1 Diagram of strategy classification (MFCC: Mel-frequency cepstral coefficient; GMM: Gaussian mixture model; SVM: support vector machine)

models with great similarity based on several distance bases, which is described in Sect. 3 in detail. Then, the Fisher score vectors with respect to the new model which can carry more discriminative information are used to train SVM. The classification information is finally obtained.

3 Feature reduction based on Gaussian mixture clustering

As described in Sect. 2, since the feature vector is based on the combination of multiple models, feature dimension will be larger with the increasing of classes and will lead to intensive computation. One solution to this problem is to reduce mixtures of the total model. If the number of mixtures used is very low, the computational complexity and memory requirement can decrease obviously.

To realize this aim, we can select some mixtures of the total model randomly. Some randomness exists when different selections of mixtures can result in different classification correct rates. We cannot know which mixtures are the best for classification.

In this paper, we examine an approach in which mixtures of the total number are selected to compose the new mixtures according to certain rules. Distances between each two mixtures are compared, and two mixtures of the nearest distance are combined into a new one. Then we calculate the distance between the new formed mixture and other mixtures to find another pair of mixtures of the nearest distance, and combine them again. Repeat the above process until the number of mixtures satisfies a certain threshold. Some distance measure criteria are shown in Table 1.

In Table 1, i and j represent the ith and jth mixture of models, respectively. $d(i,j)$ is the distance between the ith and jth mixture. Among the above criteria, resolution of Euclidean distance is low because it only considers the difference between means and ignores the difference between variances. Mahalanobis distance takes the influence of variance into account, but it cannot reflect the difference of variances when the means are equal. Bhattacharyya distance and K-L distance both consider the difference between means and variances, and emphasize similarity between models. However, K-L distance is asymmetric while K-L2 distance based on K-L is symmetrical. Therefore, we use K-L2 distance and Bhattacharyya distance as criteria to find the mixtures of the nearest distance. Experiment results are compared in Sect. 4. K-L2 distance is as follows:

$$d(i,j) = \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{|\boldsymbol{\Sigma}_i|}{|\boldsymbol{\Sigma}_j|} + \frac{|\boldsymbol{\Sigma}_j|}{|\boldsymbol{\Sigma}_i|}. \tag{5}$$

Then, the following algorithm is used to generate new mixtures using the distance criterion. It can reduce the feature dimension while the classification correct rate does not decrease.

Step 1 Train GMMs for each class.

Step 2 Merge all the Gaussian mixtures of each class into a signal model as mentioned in Sect. 2.3.

Step 3 Calculate the distance between each two mixtures of the new model according to the criterion of Bhattacharyya distance or K-L2 distance to find two mixtures of the nearest distance.

Step 4 Merge two mixtures of the nearest distance into one. Assume two mixtures of the nearest distance are i and j, the new mixture after merging is l, then

Table 1 Distance measure criteria

name	formula
Euclidean distance	$d(i,j) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$
Mahalanobis distance	$d(i,j) = (\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}\left(\frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2}\right)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$
Bhattacharyya distance	$d(i,j) = \frac{1}{8}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}\left(\frac{\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j}{2}\right)^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2}\ln\frac{\lvert(\boldsymbol{\Sigma}_i + \boldsymbol{\Sigma}_j)/2\rvert}{\lvert\boldsymbol{\Sigma}_i\rvert^{1/2}\lvert\boldsymbol{\Sigma}_j\rvert^{1/2}}$
Kullback-Leibler (K-L) distance	$d(i,j) = \frac{1}{2}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)^{\mathrm{T}}(\boldsymbol{\Sigma}_i^{-1} + \boldsymbol{\Sigma}_j^{-1})(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j) + \frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}_i^{-1}\boldsymbol{\Sigma}_j + \boldsymbol{\Sigma}_i\boldsymbol{\Sigma}_j^{-1} - 2\boldsymbol{I}_d)$

$$p_i = \frac{p(i)}{p(i)+p(j)}, \tag{6}$$

$$p_j = \frac{p(j)}{p(i)+p(j)}, \tag{7}$$

$$p(l) = p(i)+p(j), \tag{8}$$

$$\boldsymbol{\mu}_l = p_i\boldsymbol{\mu}_i + p_j\boldsymbol{\mu}_j, \tag{9}$$

$$\boldsymbol{\Sigma}_l = p_i(\boldsymbol{\mu}_i^{\mathrm{T}}\boldsymbol{\mu}_i + \boldsymbol{\Sigma}_i) + p_j(\boldsymbol{\mu}_j^{\mathrm{T}}\boldsymbol{\mu}_j + \boldsymbol{\Sigma}_j) - \boldsymbol{\mu}_l^{\mathrm{T}}\boldsymbol{\mu}_l. \tag{10}$$

Step 5 Calculate distance between the new mixture and other mixtures. If the number of mixtures reaches a threshold, go to Step 6, or else jump to Step 3.

Step 6 Use the new model with the merged mixtures to produce the Fisher score vectors. Then train SVM to classify the non-speech audio files.

4 Experiments and results

To evaluate the performance of the proposed method, several experiments are performed. This section explains different aspects of these trials.

In our experiments, the database used (http://www.comparisonics.com) contains about 1000 files including six classes (cow barks, bell, dog barks, horse barks, frog barks, evil laugh). The files we downloaded are in the following formats: .wav, .mp3, .au, .aif and .aiff. Each file's duration ranges from less than 1 s to about 1 min. All audio files are resampled to a uniform sampling rate of 8000 kHz in pulse code modulation (PCM) format with a single audio channel.

In the proposed system, we convert the audio files from their waveform representation into a sequence of 12 dimensional Mel-cepstral plus one log energy feature vectors and their first time derivatives. The cepstral and its time derivatives are combined into a 26-dimensional vector. Features are first extracted in 32 ms frames with 50% overlap in each of the two adjacent frames.

About 2/3 files are used to train the GMM of each class, the models can be used to classify by likelihood directly. In the proposed system, all the models are combined to a single model, Fisher score vectors are based on the new model. Then, use these new features to train SVMs. When training SVMs, we need to tune two parameters: one is the set of kernel parameters, the other is the regularization constant controlling the trade-off between the complexity of the function class and its ability to explain the data correctly. Here, we use cross validation to pick the parameters from a range of values that perform the best on some held-out data.

To evaluate the classification performance, accuracy rate is used. It is defined as the ratio between the number of correct-classified files and the number of total testing files.

We do two groups of experiments on several classification schemes for contrast.

The first group is to compare the effect of GMM-based classification using likelihood and SVM classification based on Fisher kernel. The two-class problem is discussed here because the correct rate of multi-class classification based on GMM is very low. Samples of dog barks and cow barks including 344 files are used. First, about 2/3 samples are used to train GMMs with different mixtures of 8, 16, 32, 64, 128 of each class. The files left are used for testing. GMM-based classification uses the likelihood to classify test samples directly. The two models are merged to one aimed at generating the Fisher score vectors. Then SVM is applied to classify test samples using the new features.

Table 2 is the correct rate of GMM and SVM classification based on Fisher kernel.

Table 2 Correct rate of GMM and SVM based on Fisher kernel

number of mixtures	correct rate of GMM/%	correct rate of SVM based on Fisher kernel/%
8	64.8256	96.4286
16	75.0000	99.1097
32	74.7093	97.3214
64	72.6744	97.3214
128	68.6046	97.3214

As shown in Table 2, the SVM classification based on Fisher kernel outperforms GMM classification greatly. The highest correct rate of the former is 99.1097% with the number of mixtures at 16, while the latter is 75.0000% with the number of mixtures at 16. There is a 24.1097% increase.

The second group is to compare the correct rate of SVM classification based on different models of before clustering, random mixture selection, and after clustering according to criterion of Bhattacharyya distance and K-L2 distance. This group infers six classes including all the files in the database. Because the effect is best when selecting the first 32 mixtures to generate the signal model on our database, all the models used later are based on this signal model. The mixture number of the signal model is $32\times6=192$, and the dimension is $32\times6\times26=4992$. The first experiment in this group is to generate the new Fisher score features with respect to the signal model directly. In the second experiment, generate a new model through selecting 64 mixtures of the signal model at random. We select five times and get the mean result. The third and fourth experiments infer mixture clustering. 192 mixtures of the signal model are clustered to 64 mixtures based on Bhattacharyya distance and K-L2 distance, respectively. Here we select 64 mixtures because its effect is better than others in our experiment. Finally, SVM is used based on new Fisher score vectors with regard to different models.

The classification accuracy of different systems is listed in Table 3.

Table 3 Correct rate and consumed time of different models

model	number of mixtures	correct rate/%	time/s
pre-clustering	192	77.1127	4.3910
select mixture randomly	64	76.4648	1.8430
Bhattacharyya distance	64	82.0423	1.9680
K-L2 distance	64	79.5775	1.7970

As shown in Table 3, although the dimension is decreased from 4992 to 1664, classification with the Fisher score features based on models after mixture clustering by Bhattacharyya distance and K-L2 yields better results than the other two methods. That is because mixture clustering can merge similar mixtures so as to remove much redundant information. Mean correct rate of selecting mixtures randomly is not very high in the table. Although the high correct rate is obtained when we select some mixtures randomly, we cannot know which mixtures can lead to a good result. There is randomness in this method. At the same time, classification based on the new model after clustering according to Bhattacharyya distance outperforms models based on the K-L2 distance. It can also be seen from Table 3 that Bhattacharyya distance is more suitable for dimension reduction than K-L2 distance. Besides the high correct rate using clustering mixtures, we can also see from Table 3 that the time used is less than without clustering. In Table 3, the last column is the time used in the course of clustering and classifying one audio sample. Because the time used in extracting Mel-frequency cepstral coefficient (MFCC) features is the same in all of the four classification methods, it is under consideration. That indicates the new method used in this paper is both more accurate and computationally more effective compared with traditional methods.

5 Conclusion

This paper proposes a simple and effective feature reduction method to realize multi-class classification of non-speech audio. The method uses clustering mixtures of GMMs to decrease number of mixtures and form a new GMM so as to decrease the dimension of feature in the Fisher space based on the new GMM. On the premise that the classification rate increases a little, it reduces the feature dimension and saves computation.

Acknowledgements This work was supported by the National Natural Science Foundation of China (Grant No. 60705019) and the National High Technology Research and Development Program of China (Nos. 2006AA010102 and 2007AA01Z417), NOKIA project, and the 111 Project (No. B08004).

References

1. Wold E, Blum T, Keislar D, Wheaton J. Content-based classification, search and retrieval of audio. IEEE MultiMedia, 1996, 3(3): 27–36
2. Rice S V. Audio and video retrieval based on audio content. Comparisonics. White Paper, 1998
3. Shirazi J, Ghaemmaghami S, Razzazi F. Improvements in audio classification based on sinusoidal modeling. In: Proceedings of 2008 IEEE International Conference on Multimedia and Expo. 2008, 1485–1488
4. Pan W J, Yao Y, Liu Z J, Huang W Y. Audio classification in a weighted SVM. In: Proceedings of International Symposium on Communications and Information Technologies. 2007, 468–472
5. Li X L, Du Z D, Zhang Y F. Kernel-based audio classification. In: Proceedings of 2006 International Conference on Machine Learning and Cybernetics. 2006, 3313–3316
6. Slaney M. Mixtures of probability experts for audio retrieval and indexing. In: Proceedings of 2002 IEEE International Conference on Multimedia and Expo. 2002, 1: 345–348
7. Guo G D, Li S Z. Content-based audio classification and retrieval by support vector machines. IEEE Transactions on Neural Networks, 2003, 14(1): 209–215
8. Giannakopoulos T, Pikrakis A, Theodoridis S. A multi-class audio classification method with respect to violent content in movies using Bayesian networks. In: Proceedings of IEEE the 9th Workshop on Multimedia Signal Processing. 2007, 90–93
9. Rabaoui A, Kadri H, Lachiri Z, Ellouze N. Using robust features with multi-class SVMs to classify noisy sounds. In: Proceedings of the 3rd International Symposium on Communications, Control and Signal Processing. 2008, 594–599
10. Jaakkola T, Diekhans M, Haussler D. A discriminative framework for detecting remote protein homologies. Journal of Computational Biology, 2000, 7(1–2): 95–114
11. Jaakkola T S, Haussler D. Exploiting generative models in discriminative classifiers. In: Solla S A, Leen T K, Müller K R, eds. Advances in Neural Information Processing Systems. Cambridge: MIT Press, 1999, 487–493
12. Smith N D, Gales M J F. Using SVMs to Classify Variable Length Speech Patterns. Technical Report CUED/F-INFENG/TR.412. 2001
13. Fine S, Navrátil J, Gopinath R A. A hybrid GMM/SVM approach to speaker identification. In: Proceedings of 2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. 2001, 1: 417–420
14. Chen L, Man H, Nefian A V. Face recognition based on multi-class mapping of Fisher scores. Pattern Recognition, 2005, 38(6): 799–811
15. Aran O, Akarun L. Multi-class classification strategies for Fisher scores of gesture and sign sequences. In: Proceedings of the 19th International Conference on Pattern Recognition. 2008, 1–4