LETTER

A Non-linear GMM KL and GUMI Kernel for SVM Using GMM-UBM Supervector in Home Acoustic Event Classification

Ngoc Nam BUI[†], *Nonmember*, Jin Young KIM[†a)], *Member*, *and* Tan Dat TRINH[†], *Nonmember*

SUMMARY Acoustic Event Classification (AEC) poses difficult technical challenges as a result of the complexity in capturing and processing sound data. Of the various applicable approaches, Support Vector Machine (SVM) with Gaussian Mixture Model (GMM) supervectors has been proven to obtain better solutions for such problems. In this paper, based on the multiple kernel selection model, we introduce two non-linear kernels, which are derived from the linear kernels of GMM Kullback-Leibler divergence (GMM KL) and GMM-UBM mean interval (GUMI). The proposed method improved the AEC model's accuracy from 85.58% to 90.94% within the domain of home AEC.

key words: *non-linear GMM KL, non-linear GUMI, audio event recognition, GMM supervector, kernel combination*

1. Introduction

According to A. Temko and C. Nadeu [1], audio event recognition has attracted considerable attention from researchers due to potential applications in various fields. They state that audio event recognition consists of several essential problems such as AEC, speech recognition, music/speech identification, audio retrieval, etc. In the scope of this paper, we mainly focus on the AEC that classifies acoustic events and activities of indoor environments. In an attempt to solve problems inherent to such environments, many AEC system structures were developed as had been previously described [1]–[3]. Although several modifications have been applied, the core system structure contains two basic stages: feature extraction and application of machine learning techniques [1]. The common features used in an AEC system are Mel-Frequency Cepstral Coefficients (MFCC), Perceptual Linear Prediction (PLP), Zero Crossing Rate (ZCR), etc. which were detailed by A. Temko and C. Nadeu [1]. Also in their work, they attempted to incorporate several learning approaches, such as GMM, SVM, and the Hidden Markov Model (HMM).

Among the different machine learning methods [1]–[5], the combination of SVM with GMM supervectors produced the most significant results [4]. Therefore, that is the main classification structure which is utilized in our AEC system. This combination was first proposed by W.M. Campbell et al. [5]. In this approach, the GMM KL kernel is applied on the basis of the approximation of the Kullback-Leibler measurement. However, the variances of the GMM supervectors have not been adapted to make the estimation of the KL kernel. This problem led to an enhanced version presented by W.M. Campbell [4]. Another advanced GUMI kernel is presented by Y.C. Huai, et al. [6], where both means and variances of the GMM supervector are utilized. T.D. Trinh, et al. experimented on AEC with the conventional SVM-GMM supervector model and MFCC feature [7]. They confirmed the performance of SVM-GMM supervector model in the home AEC domain with some kernels.

For SVM learning, high-dimensional feature mapping is adopted to deal with highly clustered data features. This results in a non-linear kernel-based SVM, which is a configuration widely implemented for pattern recognition problems [5], [8]. In particular, as explained by S. Theodoridis and K. Koutroumbas [8], the linear separability of data is an essential condition to guarantee the SVM classifier performance. They also pointed out that the features are more likely to be linear and separable when projected into a higher- or even an infinite-dimensional space. Then, a non-linear kernel is utilized to represent the inner products of the mapped data points [8]. However, the GMM KL and the GUMI are both linear kernels, and they achieved the highest rate in our AEC system [7]. Therefore, we have great interest in merging the GMM KL and GUMI with other non-linear kernels.

In this paper we introduce advanced kernels of non-linear GMM KL and non-linear GUMI. Our proposed kernels are developed using the research of M. Gönen and E. Alpaydın [9] as a basis that provides a basic structure for adapting multiple kernels. The proposed kernels combine the benefits of the means, variances, and of the non-linear kernel property into extended kernels. The evaluation of the proposed method is processed by taking an approach that is similar to a conventional SVM-GMM supervector and database as used by T.D. Trinh et al. [7].

2. GMM Supervector-Based SVM for Acoustic Event Classification

The overall structure of the AEC based on the SVM and GMM supervector mentioned by T.D. Trinh et al. [7] is described further in Fig. 1.

a) At first, the MFCC features are extracted from the audio data. Of the various techniques that are applicable, MFCC was suggested to offer one of the best feature ex-

Manuscript received November 11, 2013.
Manuscript revised April 4, 2014.

[†]The authors are with the Dept. of ECE, Chonnam National University, Gwangju, 500-757, South Korea.
a) E-mail: beyond@jnu.ac.kr (Corresponding author)

DOI: 10.1587/transfun.E97.A.1791

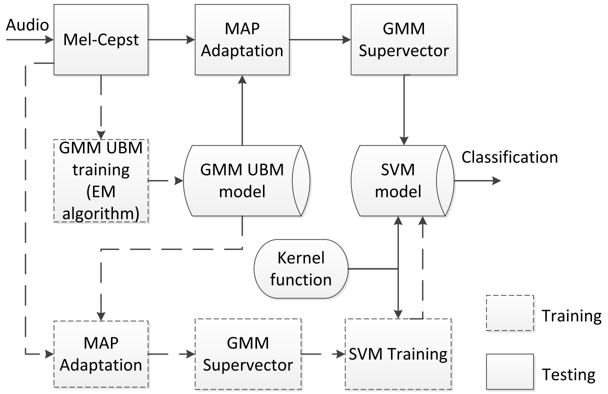

Fig. 1 The block diagram of the acoustic event classification system.

traction methods [10], [11]. Given a signal divided by a 25 ms overlapping Hamming window, the MFCC features are extracted, including the log sum power, delta, and delta-delta parameters.

b) In the next stage we utilize the GMM-UBM method. All features are pooled to train the GMM-UBM model by utilizing an Expectation Maximization (EM) algorithm.
c) After that, a GMM for each acoustic sample is derived by using a Maximum a Posterior (MAP) adaption with the UBM model.
d) A mean supervector is derived from the GMM model by cascading all Gaussian components into a single vector.
e) Finally, the SVM classifier is trained with different kernel functions.

To cope with the multiclass problem of the AEC, we adopt an SVM with a one-versus-the-rest technique, which treats the entire class as a single group and the remaining as another group. In addition, the kernel K is designed to be either linear or non-linear in order to improve the evaluation of the data. Various types of kernels have been developed to be compatible with SVM. However, we only adopt five representative implementations due to their efficiency and widespread usage, as described in prior literature [1], [3], [6], [11].

In particular, the applied kernels are a linear kernel, a 3rd polynomial kernel, an RBF kernel with $\sigma = 1$, a GMM KL divergence kernel, and a GUMI kernel. The description of the GMM KL presented by W.M. Campbell [4] is

$$K(\mathbf{g}^a \| \mathbf{g}^b) = \sum_{i=1}^{M} \left(\sqrt{\omega_i^u} (\mathbf{\Sigma}_i^u)^{-\frac{1}{2}} \boldsymbol{\mu}_i^a \right)^T \left(\sqrt{\omega_i^u} (\mathbf{\Sigma}_i^u)^{-\frac{1}{2}} \boldsymbol{\mu}_i^b \right), \qquad (1)$$

where ω_i^u, $\mathbf{\Sigma}_i^u$ are the weight and variance of the i-th component in the UBM supervector and $\boldsymbol{\mu}_i^a$ is the mean of the i-th component in the GMM supervector $\mathbf{g}^a$. The GUMI formula is described as follows [6]:

$$K(\mathbf{g}^a \| \mathbf{g}^b) = \sum_{i=1}^{M} \left[\frac{(\mathbf{\Sigma}_i^u + \mathbf{\Sigma}_i^a)^{-\frac{1}{2}}}{2} (\boldsymbol{\mu}_i^a - \boldsymbol{\mu}_i^u) \right]^T \left[\frac{(\mathbf{\Sigma}_i^u + \mathbf{\Sigma}_i^b)^{-\frac{1}{2}}}{2} (\boldsymbol{\mu}_i^b - \boldsymbol{\mu}_i^u) \right], \qquad (2)$$

where $\mathbf{\Sigma}_i^a$ and $\mathbf{\Sigma}_i^b$ are the variance of the i-th component in $\mathbf{g}^a$ and $\mathbf{g}^b$, respectively.

3. Non-linear GMM KL and GUMI Kernel

3.1 Multiple Kernel-Based SVM

A general framework for selecting multiple kernels was introduced by taking advantage of the kernel combination benefits [9]. Such a method was proven to have better performance than a single-kernel method. The definition of multiple kernels is given as:

$$k_\eta(\mathbf{g}^a \| \mathbf{g}^b) = f_\eta \left(\left\{ \eta_i k_i(\mathbf{g}_i^a, \mathbf{g}_i^b \right\}_{i=1}^{M} \right), \qquad (3)$$

where k_η is the combinatorial kernel made from different components k_i. And the kernel elements are combined using a function f_η, which can be linear or non-linear. Also, M is the number of components. The subsets $\mathbf{g}_i^a$, $\mathbf{g}_i^b$, extracted from $\mathbf{g}^a$, $\mathbf{g}^b$ feature vectors, are the parameters of the kernel k_i in turn. The weight factor η_i represents the contribution of k_i to the output kernels. When f_η is a linear combination of the kernel functions, the multiple kernel integration can be represented by

$$k_\eta(\mathbf{g}^a \| \mathbf{g}^b) = \sum_{i=1}^{M} \eta_i k_i(\mathbf{g}_i^a, \mathbf{g}_i^b). \qquad (4)$$

3.2 The Proposed Kernels

With respect to a conventional approach, the linear GMM KL and GUMI kernels do not generally follow the concept of projecting to a higher-dimensional space. Therefore, in this research, we treat the GMM KL and GUMI as combined kernels that allow merging the non-linear factor and the GMM parameters. To clarify, we first observe the GMM KL from the perspective of a projection, as described by M. Ben et al. [12]. According to them, the mean vector is projected into parameter space corresponding to each GMM component. Then the summation operator is imposed to acquire the KL distance. Following this idea, we considered the KL kernel to be a linear combination of multiple linear kernels in parameter spaces in order to further enhance the model. This combination can be expressed by the correspondence between Eq. (1) and Eq. (4) as follows:

$$k_i(\mathbf{g}_i^a, \mathbf{g}_i^b) = \left(\sqrt{\omega_i^u} (\mathbf{\Sigma}_i^u)^{-\frac{1}{2}} \boldsymbol{\mu}_i^a \right)^T \left(\sqrt{\omega_i^u} (\mathbf{\Sigma}_i^u)^{-\frac{1}{2}} \boldsymbol{\mu}_i^b \right)$$
$$\eta_i = 1 \qquad (5)$$

Eq. (5) implies that Eq. (1) is similar to Eq. (4) by replacing all GMM and UBM parameters. That is, the KL divergence kernel can be justified as a special case of the multiple kernel selection. Also, $\eta_i = 1$ indicates that all components have a similar impact on the combined output. However, the SVM with a simple, linear kernel does not function well with the overlapping data. Therefore, the points in parameter space are projected into a higher dimension using a non-linear kernel. In particular, k_i is modified by adopting a 3rd order polynomial and RBF to result in

$$k_i^{KLpol}(\mathbf{g}_i^a, \mathbf{g}_i^b) = \left((\mathbf{P}_i^a)^T \mathbf{P}_i^b + 1\right)^3 \tag{6}$$

$$k_i^{KLRBF}(\mathbf{g}_i^a, \mathbf{g}_i^b) = e^{-\frac{\|\mathbf{P}_i^a - \mathbf{P}_i^b\|_2^2}{2\sigma^2}}, \tag{7}$$

where $\mathbf{P}_i^a = \sqrt{\omega_i^u}(\boldsymbol{\Sigma}_i^u)^{-\frac{1}{2}}\boldsymbol{\mu}_i^a$ and $\mathbf{P}_i^b = \sqrt{\omega_i^u}(\boldsymbol{\Sigma}_i^u)^{-\frac{1}{2}}\boldsymbol{\mu}_i^b$. Then the result of the non-linear GMM KL kernel will produce the following:

$$k^{KLpol}(\mathbf{g}^a \| \mathbf{g}^b) = \sum_{i=1}^{M} \left((\mathbf{P}_i^a)^T \mathbf{P}_i^b + 1\right)^3 \tag{8}$$

$$k^{KLRBF}(\mathbf{g}^a \| \mathbf{g}^b) = \sum_{i=1}^{M} e^{-\frac{\|\mathbf{P}_i^a - \mathbf{P}_i^b\|_2^2}{2\sigma^2}}, \tag{9}$$

where k^{KLpol}, k^{KLRBF} indicate the KL divergence kernel with the 3rd order polynomial and RBF kernel, respectively.

Figure 2 provides a visualization of our proposed approach. At first, the GMM components are projected onto the parameter spaces in a manner similar to that described by M. Ben et al. [12]. However, the origins are still maintained. In that study, M. Ben et al. extended the approximation of the Kullback-Leibler divergence presented by M.N. Do [13] by adapting the UBM model with a Euclidean distance for the GMM supervectors. Furthermore, they also explained that the extension can be observed as a projection of the GMM supervectors into parameter space.

To extend their work, we consider that the term of $\mathbf{P}_i = \sqrt{\omega_i^u}(\boldsymbol{\Sigma}_i^u)^{-\frac{1}{2}}\boldsymbol{\mu}_i$, $i = 1, 2, \ldots, M$ is a projection of each Gaussian component into parameter space. For convenience, the GMM model in Fig. 2 only has two components and the mean vector is assumed to be in 2-D space. After that, we applied the second projection that maps the parameter space samples to higher dimensions using a non-linear kernel. This improves the linear separability of our data and allows the SVM to locate the hyperplane correctly. Then these kernel components are integrated using a linear combination.

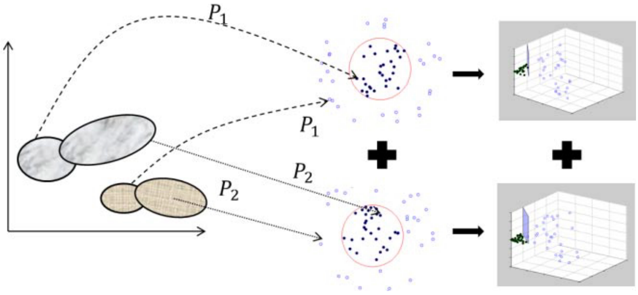

Fig. 2 Process of Mapping the GMM mean vectors to high-dimensional parameter space using a non-linear kernel.

A similar approach is undertaken for the non-linear GUMI that is normalized using the GMM supervectors with the UBM parameters. The outcomes of this approach are shown in Eq. (10) and Eq. (11).

$$k^{GUMIpol}(\mathbf{g}^a \| \mathbf{g}^b) = \sum_{i=1}^{M} \left((\mathbf{S}_i^a)^T \mathbf{S}_i^b + 1\right)^3 \tag{10}$$

$$k^{GUMIRBF}(\mathbf{g}^a \| \mathbf{g}^b) = \sum_{i=1}^{M} e^{-\frac{\|\mathbf{S}_i^a - \mathbf{S}_i^b\|_2^2}{2\sigma^2}}, \tag{11}$$

where $\mathbf{S}_i^a = \frac{(\boldsymbol{\Sigma}_i^u + \boldsymbol{\Sigma}_i^a)^{-\frac{1}{2}}}{2}(\boldsymbol{\mu}_i^a - \boldsymbol{\mu}_i^u)$ and $\mathbf{S}_i^b = \frac{(\boldsymbol{\Sigma}_i^u + \boldsymbol{\Sigma}_i^b)^{-\frac{1}{2}}}{2}(\boldsymbol{\mu}_i^b - \boldsymbol{\mu}_i^u)$.

4. Experimental Results

A same data set and model as that mentioned by T.D. Trinh et al. [7] is used to evaluate the proposed kernels. Data collection was conducted in real world environments to ensure the realistic event sounds. In particular, 41 apartments were chosen to be representative of our target environments, and data was obtained with the Zoom H2n recording equipment. The device was configured to sample signals at 16 kHz with 16 bit quantization. In this experiment, the sound pressure levels are estimated to be from 60 dB to about 90 dB, with background noise averaging about 47 dB. There are 36 groups consisting of 2,386 events. The recorded data is of about 337 MB in size. The samples are then grouped with respect to the activities being undertaken without regard for to their similarity with respect to tones. Some of the recordings, such as those that consisted of background noise or a drawer rolling sound, is even difficult to be identified by a human ear. Moreover, the intra-class variances of some of the groups are significantly high.

For example, a cell phone ringer includes various types of sounds that can be either ringing tones or short soundtracks. While most groups have more than 20 samples, some groups only have a few samples, such as that of a baby cough that is only represented in five samples. The performance degradation of the method could occur as a result of the absence of training data. Therefore, the minority groups in our dataset were discarded to maintain database quality. Finally, the remaining data consisted of 2,739 audio events, distributed unequally into 22 clusters, as can be seen in Table 1.

The 80% data were randomly selected for training and the remaining 20% were used to verify the accuracy of our method. In particular, the training set contains 2,191 samples and the test set includes the 548 remaining samples. The data selection process was repeated three different times and the accuracy rates were averaged out to ensure stability. The width factor σ in RBF kernel was chosen to be equal to 1 in all of our experiments. Finally, the accurate rates for the different kernels with varying numbers are show in Table 2, where the classification accuracy can be acquired by using Eq. (12).

Table 1 Audio sample description.

Events	Number of samples	Events	Number of samples
Baby Crying Sound	172	Front Door Bell Sound	85
Broken Glass Sound	70	Front Door Knob	78
Background Noise Sound	80	Front Door Knock	79
Boiling Water Kettle	88	Human Talking Voice	400
Cell Phone Bell	46	Screaming Sound	44
Dog Barking Sound	234	Stuff Falling Sound	56
Door Closing Sound	315	Telephone Ring Sound	40
Dish Falling Sound	89	Toilet Flushing Sound	103
Dish Friction Sound	22	Verandah Door Slide	70
Drawer Rolling Sound	68	Walking Foot Sound	498
Emergency Alarm Sound	61	Wash Stand Water	41

Table 2 Recognition rate of GMM-UBM-SVM model using conventional and other proposed kernels.

Kernel functions	Recognition rate (%)	
	16 mixtures	32 mixtures
Linear	81.11	82.02
3rd polynomial	80.57	81.72
RBF	84.40	84.42
GMM KL	84.30	84.42
GUMI	85.03	85.58
GMM KL with 3rd polynomial	84.06	84.31
GUMI with 3rd polynomial	84.61	84.36
GMM KL with RBF	**88.99**	**90.75**
GUMI with RBF	**89.36**	**90.94**

$$\text{Accuracy rate} = \frac{\text{number of correctly classified samples}}{\text{total number of testing samples}} \cdot 100\% \qquad (12)$$

The experimental results show that the proposed approach significantly improves the performance of the AEC using the SVM-GMM supervector. The classification rates were enhanced by about 5% for the 16 and 32 mixture models when the RBF kernel with GUMI or GMMKL was applied. The highest recognition rate was of about 90.94%, and it belonged to the GUMI with the RBF kernel at 32 mixtures. The integration of the GMM KL and GUMI with the non-linear kernels was effective in boosting the performance of the system dramatically.

5. Conclusion

This study proposed the use of non-linear kernels based on the multiple kernel selection model. The proposed kernels are hybrids consisting of a combination of linear and non-linear kernels. We combined the GMM KL and GUMI kernels with RBF and the polynomial kernels. The efficacy of the proposed kernels was evaluated in the domain of home AEC. In the future, more kernel combinations with varying weights will be examined to evaluate and improve the proposed approach.

Acknowledgments

This research was financially supported by the Ministry of Education, Science & Technology (MEST) and by the National Research Foundation of Korea (NRF) through the Human Resource Training Project for Regional Innovation.

References

[1] A. Temko and C. Nadeu, “Classification of acoustic events using SVM-based clustering schemes,” J. Pattern Recognition Society, vol.39, no.4, pp.682–694, April 2006.

[2] A. Temko and C. Nadeu, “Acoustic event detection in meeting-room environments,” Pattern Recognit. Lett., vol.30, no.14, pp.1281–1288, Oct. 2009.

[3] X. Zhuang, X. Zhou, M.A. Hasegawa-Johnson, and T.S. Huang, “Real-world acoustic event detection,” Pattern Recognit. Lett., vol.31, no.12, pp.1543–1551, Sept. 2010.

[4] W.M. Campbell, “A covariance kernel for svm language recognition,” Proc. ICASSP 2008, pp.4141–4144, April 2008.

[5] W.M. Campbell, D.E. Sturim, and D.A. Reynolds, “Support vector machines using GMM supervectors for speaker verification,” IEEE Signal Process. Lett., vol.13, no.5, pp.308–311, May 2006.

[6] Y. Chang Huai, L. Kong-Aik, and L. Haizhou, “An SVM kernel with GMM-supervector based on the bhattacharyya distance for speaker recognition,” IEEE Signal Process. Lett., vol.16, no.1, pp.49–52, Jan. 2009.

[7] T.D. Trinh, N.N. Bui, and J.Y. Kim, “Audio event classification using SVM with GMM-UBM supervectors,” Journal of KIIT, vol.11, no.11, pp.91–98, Nov. 2013.

[8] S. Theodoridis and K. Koutroumbas, Support Vector Machines: The Non-linear Case, Pattern Recognition, Fourth ed., pp.198–200, Academic Press, 2008.

[9] M. Gönen and E. Alpaydın, “Multiple kernel learning algorithms,” J. Mach. Learn. Res., vol.12, pp.2211–2268, July 2011.

[10] J. Breebaart and M. McKinney, “Features for audio and music classification,” Proc. ISMIR 2003, pp.151–158, Oct. 2003.

[11] T. Kinnunen and H. Li, “An overview of text-independent speaker recognition: From features to supervectors,” Speech Commun., vol.52, no.1, pp.12–40, Jan. 2010.

[12] M. Ben, M. Bester, F. Bimbot, and G. Gravier, “Speaker diarization using bottom-up clustering based on a parameter-derived distance between adapted GMMs,” Proc. ICSLP 2004, pp.2329–2332, Oct. 2004.

[13] M.N. Do, “Fast approximation of Kullback-Leibler distance for dependence trees and hidden Markov models,” IEEE Signal Process. Lett., vol.10, no.4, pp.115–118, April 2003.