GENERALIZED GAUSSIAN DISTRIBUTION KULLBACK-LEIBLER KERNEL FOR ROBUST SOUND EVENT RECOGNITION

Tran Huy Dat, Ng Wen Zheng Terence, Jonathan William Dennis, Leng Yi Ren

Human Language Technology Department, Institute for Infocomm Research, Singapore

{hdtran,wntz,studentjwd,yrleng }@i2r.a-star.edu.sg

ABSTRACT

In previous works, we have developed a spectrogram image feature extraction framework for robust sound event recognition. The basic idea here is to extract useful information from the 2D time-frequency representation of the sound signal to build up specifi feature extractions and classifier under noisy conditions. In this paper, we propose a novel robust spectrogram image method where the key is the observed sparsity of the sound spectrogram image in wavelet representations, which is modeled by the Generalized Gaussian Distributions modeling. Furthermore, the Generalized Gaussian Distribution Kullback-Leibler (GGD-KL) kernel SVM is developed to embed the given probabilistic distance into the quadratic programming machine to optimize the classification The experimental result shows the superiority of the proposed method to the previous works and the state-of-the-art in the field

Index Terms: Sound Event Recognition, Generalized Gaussian Distribution, Kullback-Leiber Distance, Kernel, Spectrogram, Wavelet

1. INTRODUCTION

Sound Event Recognition (SER) is the task to automatically detect and classify real life events using relevant information extracted from the acoustic signal. This has a wide range of applications such as acoustic surveillance [1], audio-content retrieval [2], rich transcription ASR [3], and environment understanding [4]. Unlike ASR, where the noise and environment effects can be limited using the close-talk microphones, SER always has to deal with noises, reverberation and attenuations caused by distant-microphone effects.

The motivation behind our image-based approaches comes from the auditory image concept, i.e. the visual perception of sounds through spectrogram images. It is well known that humans can easily locate the characteristic elements in a spectrogram – an approach called "spectrogram reading" [5] – and discriminate between sounds based solely on the visual information. For example, Fig. 1 shows the spectrogram for a bottle sound in both clean and noisy conditions. Here, the most characteristic, high-power elements can be easily located, even in 0dB speech babble noise as in Fig. 1b. For this and many other sound events, we have observed that the frequency spectrum is typically sparse, with the power concentrated in particular frequency bands. Therefore it is still possible to visually see the sound event among the diffuse background noise.

Our previous work on the spectrogram image feature (SIF) [6] developed a global descriptor of the pseudo-colored sound spectrogram image through the pixel distribution information. This was combined with an SVM classifie , and shows relatively good results of SER under noisy and mismatched conditions. Greater improvements were obtained when adopting a missing feature framework [7]-[8]. However, these missing feature systems requires relatively high computational cost which which may introduce difficultie in their practical implementation.

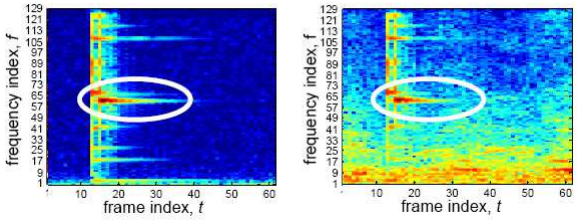

Fig. 1. Spectrogram images of bottle sound in a) clean; b) 0dB babble noise

In this paper, we develop a novel spectrogram image method for robust SER method which is effective in the evaluation and suitable for the practical implementation. The key idea of the method is the sparsity of the sound spectrogram image in wavelet domain which is nicely characterized by the Generalized Gaussian Distribution modeling. Furthermore, a novel Generalized Gaussian Distribution Kullback-Leibler (GGD-KL) kernel classificatio method is developed by embedding into the classifie the GGD-KL distance of the sound spectrogram images in their wavelet representations. Unlike our previous works [9]-[10], where the generalized Gamma distribution is used for modeling the non-negative subband power spectrum, here we employ the generalized Gaussian model which better models the wavelet coefficient [11]. An important point of our method is that the wavelet distribution is modeled for the subset of the dominating components which are extremely robust under noisy conditions. The experimental result shows the superiority of the proposed method to both the state-of-the-art and our previous methods.

We note that, the sparsity of audio spectrogram has also been studied in literature as another different framework of sparse coding [12], where a comprehensive data is used to learn the dictionary and matching pursuit is applied in the classification While achieving good performances, the sparse coding method requires extremely high computational cost and therefore is still difficul to implement in real life applications.

2. SPECTROGRAM IMAGE WAVELET REPRESENTATION

This section introduces the Spectrogram Image Wavelet Representation (SIWR). The motivation of using wavelet to represent the

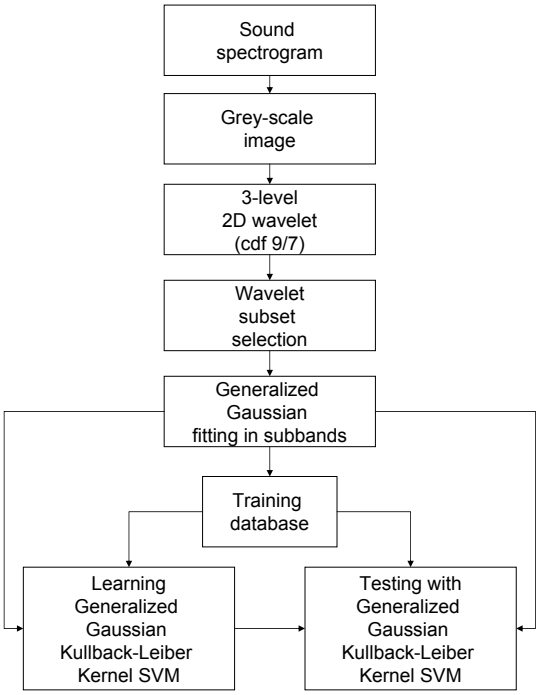

Fig. 2. Block diagram

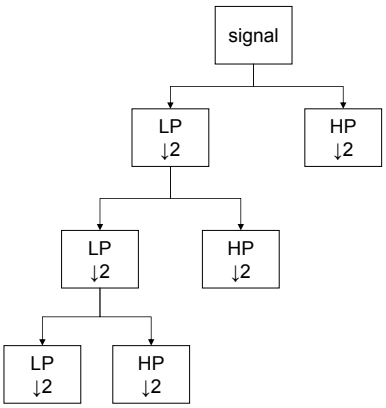

Fig. 3. Multi level wavelet transform

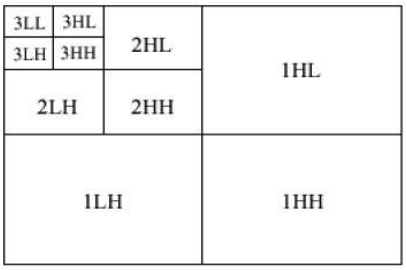

Fig. 4. Multi level 2D wavelet transform

spectrogram images comes from the observation of that the sound event spectrogram images are geometrically localized. Hence, the wavelets, which are transient, are suitable to characterize the sound event spectrograms, which are highly non-stationary. The sparsity of the wavelets is also important to deliver a robust representation in noisy conditions.

2.1. Spectrogram Image Conversion

The spectrogram image conversion follows standard signal processing algorithms. First, the audio signal is segmented into short time half overlapping windows of 20ms, before the windowed Discrete Fourier Transform is taken to transform the waveform into its spectrum, which is given by:

$$X(t,k) = \sum_{n=0}^{N-1} x(n)\omega(n)e^{-\frac{2\pi i}{N}kn} \qquad k = 0,\ldots,N-1 \quad (1)$$

where N is the length of each segment, $\omega(n)$ is the Hamming window function and k corresponds to the frequency $f(k) = \frac{kf_s}{N}$, where f_s is the sampling frequency in Hertz.

Then the log-power spectrum is given by compressing (logarithm) the power spectrum to get the spectrogram as

$$S(t,f) = 20\log_{10}|X(t,f)|. \quad (2)$$

The spectrogram matrix is then transformed into a grey-scale intensity image, with the range scaled between [0, 1]:

$$I(t,k) = \frac{S(t,k) - \min(S(t,k))}{\max(S(t,k)) - \min(S(t,k))}. \quad (3)$$

2.2. Spectrogram Image Wavelet Representation

After the sound spectrogram is converted into an grey-scale image, the wavelet transform is carried out to transform it into 2D wavelet representations. The 2D wavelet is based on 2D separable filters therefore the transform is naturally 1D and the 2D output is obtained by applying the 1D filter step-by-step in the vertical and horizontal directions. The basic idea of a 1D wavelet is illustrated in Fig. 3. In each level, the transform consists of a low-pass and a high-pass filte , which much be quadrature mirrors [12]. The multi-level transform is carried out using the same set of filter to get a wavelet tree representation, like in Fig. 3. The filte is characterized by the wavelet and scaling functions [12].

In this paper, we employ the bi-orthogonal Cohen-Daubechies-Feauvean wavelets (CDF 9/7) [13], which has been adopted in JPEG 2000 and FBI standards of image and fingerprin compressions, respectively [13]-[14]. The lifting scheme [13] is chosen in the filter implementation due to its low complexity. We note that the choice of wavelet type is not crucial and mostly based on the low cost of the implementation. The 3-level 2D wavelet transform is illustrated in Fig. 4. Each spectrogram image is transformed into 16 subbands of wavelet coefficients

$$I \xrightarrow{DWT2} \{c_k(1:n_k)\}, \quad (4)$$

where $k = 1:(L+1)^2$, $n_k = \frac{n_I}{2^k}$. L denotes the decomposition level.

2.3. Generalized Gaussian Distribution model of the SIWR

In contrast to the previous work [9]-[10], where the stochastic modeling has been applied in each subband of the power spectrum, in this paper, we model the 2D spectrogram image with the GGD model which nicely captures the sparsity of its wavelet components.

To reduce possible scaling effects, we subtract the mean of wavelet coefficient in each subband. Their distribution is now denoted by a symmetric generalized Gaussian pdf function as:

$$p(x) = \frac{\beta}{2\alpha\Gamma\left(\frac{1}{\beta}\right)} \exp\left[-\left(\frac{x}{\alpha}\right)^{\beta}\right] \tag{5}$$

where the generalized Gaussian distribution is characterized by two parameters α and β, which can be estimated using iterative methods [11]-[15].

In this paper, we derive an empirical solution for the parameter estimation based on the moment matching method, denoted by:

$$M(\beta) = \frac{E(|X|)^2}{E(X^2)} = \frac{\Gamma\left(\frac{2}{\beta}\right)^2}{\Gamma\left(\frac{1}{\beta}\right)\Gamma\left(\frac{3}{\beta}\right)}, \tag{6}$$

$$E\left(X^2\right) = \frac{\alpha^2\Gamma\left(\frac{3}{\beta}\right)}{\Gamma\left(\frac{1}{\beta}\right)}, \tag{7}$$

and utilising the following approximation of the Gamma function as [16]:

$$\log\Gamma(z) \approx (z-1) - \frac{1}{2}(z-1)^2 + \frac{1}{3}(z-1)^3. \tag{8}$$

2.4. Subset selection for modeling

To reduce the effect of noises, we choose to represent only the dominating subset of wavelet coefficient in each subband. This approach is similar to the wavelet denoising methods based on thresholding the wavelet coefficient in subbands. In this paper, we simply choose to model only the coefficient which are larger than the median value.

3. GENERALIZED GAUSSIAN DISTRIBUTION KULLBACK-LEIBER KERNEL SVM

In this section, we construct the GGD-KL kernel SVM classification to benefi from the distribution modeling of the spectrogram image wavelet representations.

3.1. SVM

Starting with linear SVM, which considers the problem of designing a separating hyperplane for m vectors $\mathbf{x}_i \in \mathbf{R^n}\ i = 1, 2, ..., m$. Each point $\mathbf{x}_i \in \mathbf{R^n}$ belongs to one of two classes, by its label $y_i \in \{1, -1\}\ , i = 1, 2, ..., m$. The goal of linear support vector machines is to fin an optimal separating hyperplane $f(\mathbf{x}) = \mathbf{w^T x} + b$, which maximizes the margin, i.e.$\frac{2}{\|\mathbf{w}^2\|}$,or equivalently minimizes $\left\|\mathbf{w}^2\right\|$

$$\begin{array}{l} \min\limits_{(\mathbf{w},\mathbf{b},\xi)} \left(\frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum\limits_{i=1}^{n}\xi_i\right) \\ \text{s.t. } y_i\left(\mathbf{w}^T\mathbf{x}_i + b\right) \geq 1 - \xi_i\ ; i = 1, 2, ..., n \\ \xi_i \geq 0, \end{array} \tag{9}$$

where the term $\sum\limits_{i=1}^{n}\xi_i$ denotes the upper bound of the misclassifica tion from the training samples and C is a coefficien that regulates between the misclassificatio and the robustness of the classificatio (width of margin). There are several ways to solve the optimization problem (9) which returns the solution in a unique form noted as

$$\mathbf{w^T} = \sum_{i=1}^{m}\alpha_i\mathbf{x}_i^T; \tag{10}$$

$$b = \sum_{i=1}^{m}\alpha_i y_i, \tag{11}$$

where $\alpha : \frac{\alpha_i}{C} = \xi_i$ is called support vector.

The separating hyperplane can be denoted now in terms of the inner product of vectors $\mathbf{x_i}$

$$f(\mathbf{x}) = \sum_{i=1}^{m}\alpha_i\mathbf{x}_i^T\mathbf{x} + b \tag{12}$$

The linear SVM can be generalized into non-linear by replacing the inner product in (12) by a kernel function

$$\mathbf{x_i^T x} \rightarrow K(\mathbf{x}, \mathbf{x_i}). \tag{13}$$

The non-linear separating hyperplane can be denoted by

$$f(\mathbf{x}) = \sum_{i=1}^{m}\alpha_i K(\mathbf{x}, \mathbf{x}_i) + b. \tag{14}$$

3.2. Probabilistic distance SVM

The deterministic non-linear kernel does not have physical meaning and in a many cases does not perform well in the classificatio tasks. Furthermore, the traditional SVM requires the input feature vectors to have the same length to conduct the kernel calculation, which create problems for preparing the samples in practice. In [9], we propose to use parametric probabilistic distance as a embedded kernel for the SVM classification i.e.

$$K(\mathbf{x}, \mathbf{x_i}) = \langle d(\mathbf{x}, \mathbf{x_i})\rangle, \tag{15}$$

where $d(\mathbf{x}, \mathbf{x_i})$ is a parametric probabilistic distance between the distributions of samples $\mathbf{x}$ and $\mathbf{x_i}$.

3.3. Generalized Gaussian Kullback-Leiber Kernel SVM

In this paper, given the modeled generalized Gaussian distribution of subband wavelet coefficients the kernel in (15) can be conducted by a closed form solution using the symmetric Kullback-Leiber distance [11], denoted as:

$$d(a, b|a_i, b_i) = \left(\frac{b_i}{b}\right)^{a_i}\frac{\Gamma\left(\frac{a_i+1}{a}\right)}{\Gamma\left(\frac{1}{a}\right)} + \left(\frac{b}{b_i}\right)^{a}\frac{\Gamma\left(\frac{a+1}{a_i}\right)}{\Gamma\left(\frac{1}{a_i}\right)} - \frac{1}{a} - \frac{1}{a_i}, \tag{16}$$

where a, b ; a_i, b_i are estimated-from-sample generalized Gaussian distribution parameters in a subband. In the training, given the kernel matrix, a proximal probabilistic distance SVM parameters can be estimated in a closed form solution [9]. In the test, for each binary pair, the decision of label y is made by evaluating (11), written as:

$$y = sign\left(\sum_{i=1}^{m} \alpha_i K\left(\mathbf{x}, \mathbf{x}_i\right) + b\right). \quad (17)$$

Multi class SVM is firs performed in each subband and then summarized over subbands. In each step, the conventional maximum voting principle is applied. This scheme is found to perform better than the sum-over-subband kernel implemented in [9].

4. EXPERIMENTS

4.1. Database

The task is to classify unknown sound samples into one of the ten sound classes: normal speech, cry, scream, laugh, knock, breaking, door slamming, phone ring, explosion and clapping. The database consists of about 2 hours of audio taken from [18]. The audio samples are of approximately one-two seconds in length.

To validate the robustness of the proposed techniques, we played back and recorded the whole audio data in three environments: in an offic with SNR ranging from 20dB to 25dB; in a hall with SNR ranging from 10dB to 15 dB, and at a shopping mall with SNR ranging of 0-5 dB. In the last two cases, the reverberation also contributes to the noise in SNR calculations. We note that, the segmental SNR (Signal to Noise Ratio) is estimated by algorithm developed in an our previous paper [20].

To compare the performances, we evaluate the classificatio accuracy in 10 runs cross-validations. In each run, around one hour of testing data are randomly selected. More precisely, 2794 training and 2782 testing sample names are selected for each run.

Excepting the multi-condition HMM method, when the training samples are taken from all four conditions (i.e. clean, office hall, mall), the rest of methods are based on clean training, i.e. 2782 name-selected testing samples are from clean data while for each noisy data, 2794 name-selected training samples are taken

4.2. Evaluation methods

To validate our proposal, we firs compare the proposed method to the state-of-the-art methods which are the MFCC-HMM with clean training and the MC-MFCC-HMM with multi-condition training (mentioned above).

1. MFCC-HMM [19]: 39-dimensional frame MFCC feature with 3-stage HMM with 8-component diagonal GMM models [18].
2. MC-MFCC-HMM: the multi-condition training with the conventional MFCC-HMM framework.
3. SC-MP: the sparse coding with matching pursuit classifie method [12]

Next, we compare the proposed to our previous methods, as follows,

1. SIF-SVM: global spectrogram image descriptor with SVM classifie [6].
2. MF-MFCC-SVM: our visual inspired (blob detection) missing feature method [8].
3. HE-STE-SPDSVM: our previous Hellinger-exponential kernel SVM with subband temporal envelop representations [10].

Table 1. Comparison to state-of-art methods

Conditions	MFCC-HMM	MC-MFCC-HMM	SC-MP
Offic	90.1 ± 2.5	91.2 ± 2.1	94.0 ± 1.4
Hall	58.1 ± 2.9	81.4 ± 2.7	63.1 ± 3.7
Mall	38.1 ± 4.4	72.1 ± 2.2	56.1 ± 2.1

Table 2. Comparison to our previous methods

Conditions	MF-MFCC-GMM	SIF-SVM	HE-STE-SPDSVM	Proposed
Offic	90.1 ± 2.9	92.4 ± 1.9	96.4 ± 2.3	97.0 ± 2.4
Hall	82.6 ± 3.9	79.7 ± 3.5	82.5 ± 3.2	83.1 ± 2.7
Mall	73.2 ± 3.7	70.8 ± 3.5	72.4 ± 3.4	74.1 ± 3.8

4.3. Results and discussions

Table 1 compares the proposed methods to the state-of-art MFCC-HMM in clean training and multi-condition training, respectively. The results of MFCC-GMM and MFCC-SVM are slightly worse and therefore were not included in the comparison. We can clearly see that the proposed method significantl outperformed the state-of-art MFCC-HMM, which has the same clean training conditions. The proposed method even outperformed the multi-condition training, which requires much more training resources. Statistical significanc ($p < 0.1$) of proposed to the multi-condition method was obtained in two conditions (20-25dB and 10-15dB). The SC-MP has shown not effective compared to the MC-MFCC-HMM.

Next, we compare the proposed method to our previous methods. The proposed method shows a statistically significan improvement to each of the previous SIF (global descriptor of the spectrogram image), MF-MFCC-HMM (blob detection missing feature), and HE-STE-SPDSVM (Hellinger-exponential kernel with auditory-inspired subband temporal envelope representations) methods in the offic environment. In the lower SNR conditions, the proposed method outperformed SIF and has a comparable performance to the other methods. We also note that all the visual inspired methods have shown significan superiority to the state-of-art MFCC-HMM in clean training, and are comparable to the MC-MFCC-HMM with multi-condition training. Therefore, this demonstrates the benefi of using 2D information for sound event recognition.

Note that, although the proposed GGD-KL kernel is not factorisable to transform it into linear SVM as the Hellinger-exponential in [8], the evaluation of (14) is reasonably fast thanks to the sparsity of the support vectors.

5. CONCLUSIONS

We propose a novel representation of the sounds based on the Generalized Gaussian distribution model of the spectrogram image wavelet coefficients Based upon on the proposed representation, we further develop the GGD-Kullback-Leiber kernel SVM classificatio method. The method takes into account the advantages of spectrogram image, JPEG 2000 wavelet compression, wavelet denoising framework, and the embedded probabilistic distance SVM, and has shown the superiority in the noisy and mismatched conditions.

6. REFERENCES

[1] C. Clavel, T. Ehrette, and G. Richard, Event detection for an audio-based surveillance system, in IEEE ICME, Amsterdam, July 2005.

[2] E. Wold, T. Blum, D. Keislar, and J. Wheaton, Content-based classification search, and retrieval of audio, IEEE Multimedia, vol. 3, pp. 27-36, 1996.

[3] The 2007 (RT-07) Rich Transcription Meeting Recognition Evaluation Plan, http://www.nist.gov/speech/tests/rt/rt2007

[4] Jia-Ching Wang and Hsiao-Ping Lee and Jhing-Fa Wang and Cai-Bei Lin, Robust environmental sound recognition for home automation, IEEE JASE,v.5, n.1, pp.25-31, 2008

[5] Zue, V., Notes on Spectrogram Reading, Dept of EECS Technical Document, MIT, Cambridge, 1985.

[6] Jonathan Dennis, Tran H.D, and H. Li, Spectrogram image feature for sound event classificatio in mismatched conditions, IEEE Signal Process. Lett., v.18, n.2, pp. 130-133, 2011.

[7] Jonathan Dennis,Tran Huy Dat and Engsiong Chng , Image feature representation of the subband power distribution for robust sound event classification IEEE Transactions on Audio, Speech & Language Processing, v.21, n2., pp.367-377, 2013.

[8] Yi Ren Leng and Tran Huy Dat, Using blob detection in missing feature linear-frequency cepstral coefficient for robust sound event recognition, in Proc. INTERSPEECH 2012, 2012.

[9] Tran H.D., and H. Li, Sound event recognition with probabilistic distance SVMs, IEEE Transactions on Audio, Speech & Language Processing, 19(6), 1556-1568, 2012

[10] Tran H.D., and H. Li, Probabilistic distance SVM with Hellinger-Exponential kernel for sound event classification in Proc IEEE ICASSP 2011, pp.2272-2275, 2011.

[11] Do M. N., and Vetterli, M., Wavelet-based texture retrieval using generalized Gaussian density and Kullback-Leibler distance, IEEE Transactions on Image Processing, 11(2), pp.146-158, 2002.

[12] Scholler, Simon and Purwins, Hendrik, "Sparse coding for drum sound classificatio and its use as a similarity measure," Proceedings of 3rd international workshop on Machine learning and music, MML '10, 2010

[13] C. Christopoulos, A. Skodras, and T. Ebrahimi, The JPEG2000 Still Image Coding: An Overview, IEEE Transactions on Consumer Electronics, Vol. 46, No. 4, pp. 1103-1127, November 2000

[14] M. Unser and T. Blu., Mathematical properties of the JPEG2000 wavelet filter IEEE Trans. on Image Proc., vol. 12, no. 9, Sep. 2003.

[15] Brislawn, C. (2002). The FBI fingerprin image compression specification Wavelet Image and Video Compression, 271-288.

[16] Varanasi, M.K., Aazhang B., Parametric generalized Gaussian density estimation, J. Acoust. Soc. Am, v.86, pp. 1404-1415, 1989.

[17] Calculators and the Gamma Function at http://www.rskey.org/gamma.htm

[18] Sound Effect Collections at http://www.sound-ideas.com/

[19] CLEAR 2006, Southampton, UK, April 6-7, 2006, Lecture Notes in Computer Science Springer Berlin / Heidelberg, Volume 4122/2007, 978-3-540-69567-7

[20] Tran H. D., K. Takeda, F. Itakura, "On-line Gaussian mixture modeling in the log-power domain for signal-to-noise ratio estimation and speech enhancement,", Speech Communication 48(11): 1515-1527 (2006)