World Academy of Science, Engineering and Technology
International Journal of Electronics and Communication Engineering
Vol:6, No:8, 2012

Multiclass Support Vector Machines for Environmental Sounds Classification Using log-Gabor Filters

S. Souli, Z. Lachiri

***Abstract*—**In this paper we propose a robust environmental sound classification approach, based on spectrograms features driven from log-Gabor filters. This approach includes two methods. In the first methods, the spectrograms are passed through an appropriate log-Gabor filter banks and the outputs are averaged and underwent an optimal feature selection procedure based on a mutual information criteria. The second method uses the same steps but applied only to three patches extracted from each spectrogram.

To investigate the accuracy of the proposed methods, we conduct experiments using a large database containing 10 environmental sound classes. The classification results based on Multiclass Support Vector Machines show that the second method is the most efficient with an average classification accuracy of 89.62 %.

***Keywords*—**Environmental sounds, Log-Gabor filters, Spectrogram, SVM Multiclass, Visual features.

I. Introduction

THE research of environmental sound classification is less developed than that of speech and music classification. Recently, some efforts have been interested on classifying environmental sounds [1]-[3], which the objective is to offer many services, for instance surveillance and security applications.

In addition, the sound recognition systems used are based on different descriptors such as classic acoustic descriptors, cepstral descriptors, spectral descriptors, and time-frequency descriptors. These descriptors can be used as a combination of some, or even all, of these 1-D audio features together, but sometimes the combination between descriptors [2] increases the classification performance compared with the individual used features. The problem is that there are many features which negatively influence in the quality of classification. Therefore, the recognition rates decrease when the number of targeted classes' increases because of the presence of some difficulties likes randomness and high variance [1]. There is possibility of investigating the visual domain, what allowed classifying the musical sounds which based on texture images [4] and [5], and the environmental sounds in [6], where we have demonstrated that the time-frequency representation can be applied to environmental sounds and can produce a good result for classification.

S. Souli is with the Signal, Image and Information Technology Lab, ENIT BP 37, 1002, Le Belvédère, Tunisia (phone: 216-24-722-792 ; e-mail: soulisameh@yahoo.fr).

Z. Lachiri is with the Signal, Image and Information Technology Lab, ENIT BP 37, 1002, Le Belvédère, Tunisia (e-mail: zied.lachiri@enit.rnu.tn).

Besides, many studies likes [7] and [8] show that spectro-temporal modulations play an important role in sound perception, and stress recognition in speech [9], in particular the 2D Gabor, which are suitable and very efficient to feature extraction. They offer an excellent simultaneous localization of spatial and frequency information [9]. They have many useful and important properties, in particular the capacity to decompose an image into its underlying dominant spectro-temporal components. The Gabor filters represent the most effective means of packing the information space with a minimum of spread and hence a minimum of overlap between neighboring units in both space and frequency [10].

In this paper, we develop a two nonlinear feature extraction methods in the visual domain based on time frequency representation as the primary features and the log-Gabor filters were then employed to derive the secondary features. For classification, we use the SVM's with multiclass approach: One-Against-One.

This paper is organized as follows. Section II describes the visual feature extraction methods and presents the classification algorithm. Classification results are given in Section III. Finally conclusions are presented in Section IV.

II. Environmental Sound Classification System

Our classification system consists of three methods. In the first method, a spectrogram is generated from environmental sound. Next, we pass to the single log-Gabor filters extraction phase, to construct Gabor coefficients for two scales and six orientations. Then, we applied mutual information for the obtained Gabor coefficients to get the optimal feature which finally used in the classification phase. The second method is interested in the same technique as the first method, but in this case, we introduce 12 log-Gabor filters for each spectrogram instead of single log-Gabor filter. Then we passed to averaged operation flowed by a mutual information to optimize the obtained feature to facility the classification.

In the third method, we introduce the spectrogram patch notion. Our idea is to segmented spectrogram into 3 patches. Intuitively, for each patch, 12 log-Gabor filters are calculated, after that we applied an averaged operation, then a mutual information selection for passed in the classifier.

For the classification, we employ SVM, in One-Against-One configuration with the Gaussian kernel (Fig. 1).

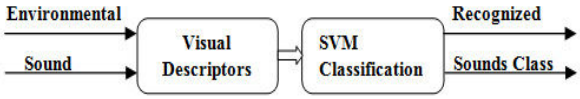

Fig. 1 Classification System Overview

A. Advantages and Calculation of Environmental Sound Spectrogram

The spectrogram is the most current time-frequency representation. It is a visual energy representation across frequencies and over time. The horizontal axis represents time, and the vertical axis is frequency [11].

With spectrogram we can observe the complete spectrum of environmental sounds and express sound by combining the merit of time and frequency domains [12]. Furthermore, we can easily identify the environmental sounds spectrograms by their contrast, since they are considered as different textures Fig. 2 [4]. These observations show that the spectrograms contain characteristics which can be used to differentiate between different environmental sounds class [13]. The sound time-frequency contains a large amount of information and provides a representation that can be easily interpreted [14]. The Short-Time Fourier Transform (STFT) was used to calculate the spectrogram $s(x,y)$, and the frames were taken to be 256-point frames with 192-point overlap.

Let $f[n]$ be an audio signal, $n = 0,1,\dots,N-1$.

The time-frequency transform factorizes f over a family of time-frequency atoms $\{g_{x,y}\}_{x,y}$ where x and y are respectively time and frequency. The short-time Fourier transform of f is defined by [15]:

$$F[x,y] = \langle f, g_{x,y} \rangle = \sum_{n=0}^{N-1} f[n]\, g^*_{x,y}[n] \quad (1)$$

where $*$ is the conjugate. The atoms of the short-time Fourier transform are:

$$g_{x,y}[n] = w[n-l] exp\left(\frac{i2\pi kn}{k}\right) \quad (2)$$

where $w[n]$ is the Hamming window, for each $0 \le y < k$, $F[x,y]$ is calculated for $0 \le y < k$.

The classification is based on the log-spectrogram:

$$s(x,y) = log|F[x,y]| \quad (3)$$

Let us take the spectrograms of environmental sounds as illustrated in Fig. 2, each class contains sounds with very different temporal or spectral characteristics, levels, duration, and time alignment for example door slams present a wide frequency band but with a short duration.

In addition, for the children voices we can distinguish the presence of the privilege frequencies. Concerning phone rings, we remark that it presents fundamental frequencies. Another remark about phone rings and children voices, they are harmonic sounds. Furthermore, we notice that there are some similarities between explosions and gunshots though, they belong to different classes. We also illustrate according to Fig. 2 that there are signals which present textural properties can be easily learned without explicit detailed analysis of the corresponding patterns [2], so easy to be distinguished, which influences in a positive way in the phase of the classification.

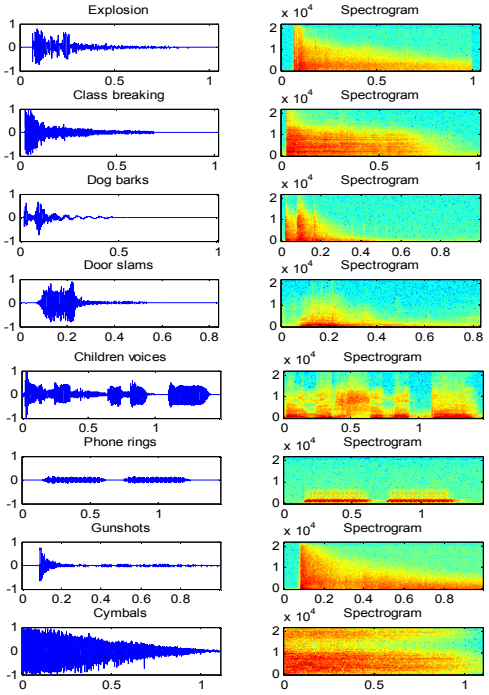

Fig. 2 Audio Waveform and Spectrograms of 8 Classes Environmental Sound

B. Log-Gabor Filters

Log-Gabor filters provide a way to extract features that can describe environmental sounds where other audio features fail. It will be shown that the log Gabor feature can be used to yield higher recognition accuracy for environmental sounds.

The log-Gabor function in the frequency domain can be described by the transfer function $G(r,\theta)$ with polar coordinates [9]:

$$G(r,\theta) = G_{radial}(r).G_{angular}(r) \quad (4)$$

where $G_{radial}(r) = e^{-\log(r/f_0)^2/2\sigma_r^2}$, is the frequency response of the radial component and $G_{angular}(r) = exp\,(-(\theta/\theta_0)^2/2\sigma_\theta^2)$ represents the frequency response of the angular filter component.

We note that (r,θ) are the polar coordinates, f_0 represents the central filter frequency, $\theta_0$is the orientation angle, σ_r and σ_θ represent the scale bandwidth and angular bandwidth respectively.

The log-Gabor feature representation $|S(x,y)|_{m,n}$ of a magnitude spectrogram $s(x,y)$ was calculated as a convolution operation performed separately for the real and imaginary part of the log-Gabor filters:

$$Re(S(x,y))_{m,n} = s(x,y) * Re\big(G(r_m,\theta_n)\big) \quad (5)$$

$$Im(S(x,y))_{m,n} = s(x,y) * Im\big(G(r_m,\theta_n)\big) \quad (6)$$

(x,y) represent the time and frequency coordinates of a spectrogram, and $m = 1,\dots,N_r = 2$ and $n = 1,\dots,N_\theta = 6$ where N_r devotes the scale number and N_θ the orientation number. This was followed by the magnitude calculation for the filter bank outputs:

$$|S(x,y)| = \sqrt{\big(Re\big(S(x,y)\big)_{m,n}\big)^2 + Im(S(x,y))_{m,n})} \quad (7)$$

The averaged operation was calculated for each 12 log-Gabor filter, appropriate for each spectrogram, which purpose is to obtain a single output array:

$$\big|\hat{S}(x,y)\big| = \frac{1}{N_r\,N_\theta}\sum_{\substack{m=1\\n=1}}^{N_r,N_\theta} |S(x,y)|_{m,n} \quad (8)$$

C. Descriptors Extraction Methods

We have applied two spectrogram-based methods to the features extraction. These methods use the log-Gabor filters as spectro-temporal component expansion.

- *12 log-Gabor Filters Concatenation*

In this approach, each environmental sound spectrogram was passed thought a bank of 12 log-Gabor filters. This produced a bank of 12 log-Gabor filters $\{G_{11}, G_{12}, \dots, G_{16}, G_{21}, \dots, G_{25}, G_{26}\}$, with each filter representing different scale and orientation; we can see Fig. 4 for more explication. Thus, this result allows us to say that we obtain for each spectrogram a bank of 12 log-Gabor filters. These resulting feature values were then concatenated into 1D-vectors. Then the computation of averaged, passed thought the MI criteria, and was sent to the SVM for classification.

Fig. 3 Descriptors extraction using 12 log-Gabor filters

- *Three Spectrogram Patches with 12 log-Gabor Filters*

In this approach, the concept is to use the spectrogram patch, the purpose is to find the suitable part of spectrogram, where concentrates the efficient structure, which gives a better result. The idea is to extract three patches from each spectrogram. In effect, the 2-D Gabor filter decomposes a patch into its spectro-temporal components [10]. The first patch included frequencies from 0.01Hz to 128Hz, the second patch, from 128Hz to 256Hz, and the third patch, from 256Hz to 512Hz.

Indeed, each patch was passed through 12 log-Gabor filters, followed by an averaged operation and then passed to MI feature selection algorithm, which constitute the parameter vector for the classification (Fig. 4).

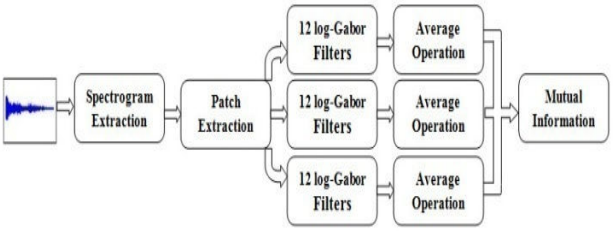

Fig. 4 Descriptors extraction using 3spectrogram patches with12 log-Gabor filters

D. Features Optimization Using Mutual Information

The information found commonly in two random variables is defined as the mutual information between two variables X and Y, and it is given as [16]:

$$I(X;Y) = \sum_{x\in X}\sum_{y\in Y} p(x,y)log\frac{p(x,y)}{p(x)p(y)} \quad (9)$$

where $p(x) = Pr(X = x)$ is the marginal probability density function and $p(x) = Pr(X = x)$, and $p(x,y) = Pr(X = x, Y = y)$ is the joint probability density function. The purpose of the mutual information is to optimize the size of feature vectors to facilitate the classification phase.

E. SVM Classification

For evaluation we use a Support Vector Machines, in a One-against-One configuration. In the nonlinear case, the idea is to use a kernel function $K(x_i, x_j)$, which satisfies the Mercer conditions, where each x_i is the feature vector of a signal. Here, we used a Gaussian RBF kernel:

$$k(x,x') = exp\left[\frac{-\|x-x'\|^2}{2\sigma^2}\right] \quad (10)$$

where $\|.\|$ indicates the Euclidean norm in $\Re^d$.

We hence adopted one-against-one approach of multiclass classification, this method consists in creating a binary classification of each possible combination of classes, and the result for k classes is $k(k-1)/2$. The classification is then carried out in accordance with the majority voting scheme [17].

III. Classification Results

Our corpus of sounds comes from commercial CDs [18]. Among the sounds of the corpus we find: explosions, broken glass, door slamming, gunshot, etc. We used 10 classes of environmental sounds as shown in Table I.

TABLE I
CLASSES OF SOUNDS AND NUMBER OF SAMPLES IN THE DATABASE USED FOR PERFORMANCE EVALUATION

Classes	Train	Test	Total
Door slams *(Ds)*	208	104	312
Explosions *(Ep)*	38	18	56
Class breaking *(Cb)*	38	18	56
Dog barks *(Db)*	32	16	48
Phone rings *(Pr)*	32	16	48
Children voices *(Cv)*	54	26	80
Gunshots *(Gs)*	150	74	224
Human screams *(Hs)*	48	24	72
Machines *(Mc)*	38	18	56
Cymbals *(Cy)*	32	16	48
Total	670	330	1000

All signals have a resolution of 16 bits and a sampling frequency of 44100 Hz that is characterized by a good temporal resolution and a wide frequency band. Most of the signals are impulsive; we took 2/3 for the training and 1/3 for the test.

We suggested for classification the cross-validation procedure. Indeed, according to [19], this method consists in setting up a grid-search for γ and C**.** For the implementation of this grid, it is necessary to proceed iteratively, by creating a couple of values γ and C. In this work, we use the following couples:

$$C, \gamma: C=[2^{(-5)}, 2^{(-4)},\ldots, 2^{(15)}] \text{ et } \gamma =[2^{(-15)}, 2^{(-14)}, \ldots, 2^{(3)}].$$

Results of the first approach and second approach are illustrated in Table II. Indeed, let us begin by the first method, which the idea consists of 12 log-Gabor filters concatenation, and then the averaged operation is applied, followed by the mutual information criteria.

The obtained classification results range from 62.50% to 99.35% for the environmental sounds data.

TABLE II
RECOGNITION RATES FOR AVERAGED OUTPUTS OF 12 LOG-GABOR FILTERS AND 3 SPECTROGRAM PATCHES APPLIED TO ONE-AGAINST-ONE SVM'S BASED CLASSIFIER WITH GAUSSIAN RBF KERNEL

Classes	12 Log-Gabor Filters		3 Spectrogram Patches	
	Kernel Parameters (C,γ)	Classif. Rate (%)	Kernel Parameters (C,γ)	Classif. Rate (%)
Ds	$(2^{(-5)},2^{(2)})$	99,35	$(2^{(-5)},2^{(-6)})$	94.87
Ep	$(2^{(-4)},2^{(-6)})$	62.50	$(2^{(-4)},2^{(-6)})$	69.64
Cb	$(2^{(-5)},2^{(-4)})$	78.57	$(2^{(-5)},2^{(-4)})$	78.57
Db	$(2^{(-3)},2^{(1)})$	87.50	$(2^{(-3)},2^{(1)})$	89.58
Pr	$(2^{(-5)},2^{(2)})$	83.33	$(2^{(-5)},2^{(-4)})$	87.50
Cv	$(2^{(-5)},2^{(2)})$	87.50	$(2^{(-5)},2^{(2)})$	82.50
Gs	$(2^{(-5)},2^{(2)})$	98.21	$(2^{(-5)},2^{(2)})$	83.03
Hs	$(2^{(-3)},2^{(1)})$	94.11	$(2^{(-3)},2^{(-4)})$	95.58
Mc	$(2^{(-3)},2^{(-10)})$	89.28	$(2^{(-3)},2^{(-10)})$	92.85
Cy	$(2^{(-5)},2^{(2)})$	95.83	$(2^{(-5)},2^{(2)})$	93.75

We were able to achieve an averaged accuracy rate of the order 89.62% in ten classes with one-against-one approach.

In the second approach results, we obtained an averaged accuracy rate of the order 86.78%, so this result is slightly lower than the first method result.

The results of the experiments are satisfactory and encouraging to investigate better the visual domain.

We remark that our system had significant results and we notice also that our visual features are significantly outperforms in spite of limited number of features. Furthermore, the used feature vector represents all relevant information in the signals to recognize.

According to the experimental results, our methods achieved enhancement performances.

The adjustment of the extraction methods of visual features, used in image processing, to the special characteristics of the environmental sounds has given satisfactory and improved classification results.

IV. CONCLUSION

In this paper, new methods for environmental sound classification are presented, based on the visual domain. These methods used spectro-temporal decomposition with log-Gabor filters.

The first method utilized an average of 12 log-Gabor filters concatenation. The second method segmented spectrogram into three patches with averaged 12 log-Gabor filters.

We show that the first method achieve an averaged accuracy of 89.62%, it obtained the best classification result compared to first and third methods. In the future work, our idea consists of more exploration in the visual domain for environmental sound classification.

REFERENCES

[1] S. Chu, S. Narayanan, and C.C.J. Kuo, "Environmental Sound Recognition with Time-Frequency Audio Features," *IEEE Trans. on Speech, Audio, and Language Processing*, vol. 17,no. 6 pp. 1142-1158, 2009.

[2] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze. "Using One-Class SVMs and Wavelets for Audio Surveillance," *IEEE Transactions on Information Forensics and Security.* Vol. 3, no.4, pp. 763-775, 2008.

[3] K. El-Maleh, A. Samouelian, and P. Kabal, "Frame-level noise classification in mobile environments," *in Proc. ICASSP, Phoenix, AZ,* 1999, pp.237–240.

[4] G.Yu, and J.J. Slotine. "Audio Classification from Time-Frequency Texture," *In Proc. IEEE. ICASSP, Taipei,* 2009, pp. 1677–1680.

[5] G. Yu, and J. J. Slotine, "Fast Wavelet-based Visual Classification," in Proc. *IEEE International Conference on Pattern Recognition ICPR, Tampa*, 2008, pp.1-5.

[6] S. Souli, Z. Lachiri, "Environmental Sounds Classification Based on Visual Features," *CIARP, Springer, Chile,* vol.7042, pp. 459-466, 2011.

[7] M. Kleinschmidt, "Methods for capturing spectro-temporal modulations in automatic speech recognition," *Electrical and Electronic Engineering Acoustics, Speech and Signal Processing Papers, Acta Acustica,* vol.88, no. 3, pp.416-422, 2002.

[8] M. Kleinschmidt, "Localized spectro-temporal features for auto-matic speech recognition," *in Proc. Eurospeech,* 2003, pp.2573-2576.

[9] L. He, M. Lech, N. Maddage, and, N. Allen,"Stress and Emotion Recognition Using Log-Gabor Filter," *Affective Computing and Intelligent Interaction and Workshops, ACII, 3rd International Conference on, Amsterdam,* 2009, pp.1-6.

[10] T. Ezzat, J. Bouvrie, and T. Poggio, "Spectro-Temporal Analysis of Speech Using 2-D Gabor Filters," *Proc. Interspeech,* Citeseer, 2007, pp. 1-4.

[11] T. Lamper, A. O'Keefe, S. E.M. "A survey of spectrogram track detection algorithms," *Applied Acoustics.* vol. 71, pp. 87–100, 2010.

[12] Z. Xinyi, Y. Jianxiao, H. Qiang. "Research of STRAIGHT Spectrogram and Difference Subspace Algorithm for Speech Recognition," *Int. Congress On Image and Signal Processing (CISP), IEEE DOI Link ,* 2009, pp.1-4.

[13] L. He, M. Lech, N. C. Maddage and N. Allen. "Stress Detection Using Speech Spectrograms and Sigma-pi Neuron Units," *int. Conf. on Natural Computation*, 2009, pp. 260-264.
[14] J. Dennis, and H.D.Tran, and H. Li. "Spectrogram Image Feature for Sound Event Classification in Mismatched Conditions," *Signal Processing Letters, IEEE*. vol. 18, pp. 130-133, 2011.
[15] G. Yu, S. Mallat and E. Bacry. "Audio Denoising by Time-Frequency Block Thresholding," *IEEE Transactions on Signal Processing,* vol 56, pp. 1830-1839, 2008.
[16] N. Kwak, C. Choi, "Input Feature Selection for Classification Problems," *IEEE Trans, On Neural Networks*, vol. 13, no.1, pp. 143-159, 2002.
[17] B. Scholkopf, and A. Smola, "*Learning with Kernels,*" MIT Press, 2001.
[18] The Leonardo Software website. [Online]. Available: http: //www.leonardosoft.com. Santa Monica, CA 90401.
[19] C.-W. Hsu, C-C. Chang, C-J. Lin, "A practical Guide to Support Vector Classification," *Department of Computer Science and Information Engineering National Taiwan University, Taipei, Taiwan,* 2009.