Audio Surveillance under Noisy Conditions using Time-Frequency Image Feature

Roneel V. Sharan and Tom J. Moir
School of Engineering
Auckland University of Technology
Private Bag 92006, Auckland 1142, New Zealand
Email: roneel.sharan@aut.ac.nz, tom.moir@aut.ac.nz

Abstract—**In this paper, we use the novel method of using features extracted from the time-frequency image representation of a sound signal in an audio surveillance application. In particular, we investigate two image representations: linear grayscale and log grayscale. We first divide a sound signal into smaller frames and apply a windowing function. The absolute value of the Discrete Fourier Transform of each frame is then computed and normalized to get the intensity values for the linear grayscale image. The generation of the log grayscale image takes a similar approach but we take log power of the values before data normalization. Each image is then divided into blocks and central moments are computed in each block. We carry out experimentation under different noise conditions and varying signal-to-noise ratio using support vector machines for classification. Based on the classification accuracy, the linear grayscale image approach is found to be more noise robust than the log grayscale image approach. It was also found to perform better than using mel-frequency cepstral coefficients as features which is a common baseline feature in most sound recognition applications.**

Keywords—audio surveillance; central moments; linear grayscale; log grayscale; signal-to-noise ratio; sound recognition; spectrogram; time-frequency image

I. INTRODUCTION

Mel-frequency cepstral coefficients (MFCCs) have been used as a baseline feature in many sound recognition systems such as in [1-3]. However, MFCCs have been shown to perform poorly in noisy conditions [4]. As such, they are often complemented with other features but the addition of new features does not necessarily increase the classification accuracy. Some features are redundant and different combination of features need to be experimented with to determine the one that performs best as seen in [3]. Optimization techniques such as genetic algorithms can also be used for this purpose as in [5].

However, in a recent study [6], features extracted from the time-frequency image, or spectrogram image, were shown to give promising results in sound event recognition under noisy conditions. Literature on this unique approach is limited though. In [7], time-frequency images of the sound signal were used for feature extraction in a hearing aid application. Eleven features were extracted for classifying four classes: *speech*, *speech in noise*, *noise*, and *classical music*. The original image is in grayscale but binary images were also created for feature extraction. Some of the features extracted are: sharpness in peak of histogram of the grayscale image, variation in frequency of the histogram, number of pixels that have 1 pixel line width across the *y* axis (frequency axis), mode value of the 256 gray levels in the histogram, maximum frequency of histogram, mean value of histogram, number of isolated points in the binary image, number of white pixels in the binary image, and power difference between low frequency and high frequency. Five features are firstly used to classify between *classical music* and *the others*. *The others* is then classified as *speech*, *speech in noise*, and *noise* using the remaining six features.

A similar approach was taken by Costa et al. [8] in music genre recognition. In their work, the audio signal is first converted to a spectrogram using time decomposition [9] and the gray level co-occurrence matrix (GLCM) [10] texture descriptors are extracted as features using a zoning technique with a total of 10 zones. The following seven descriptors were used in their work: entropy, correlation, homogeneity, third order momentum, maximum likelihood, contrast, and energy. A support vector machine (SVM) classifier with maximum voting strategy was used with three-fold cross-validation. The results are compared against those in [11] which takes an instance-based approach with feature vectors represented by short-term, low-level characteristics of the music audio signal. Only a marginal increase is seen in the average classification accuracy but results showed an improvement of 7 percentage points when the two methods were combined.

In [6], however, a slightly different approach to the ones given above was used. The spectrogram images were partitioned into 9×9 blocks and second and third central moments were computed in each block. As such, the final feature vector for the grayscale image is 162-dimensional. The results under noisy conditions were shown to outperform those using MFCCs as features.

In this paper, we propose to use the approach given in [6] for audio surveillance application in noisy conditions. However, when compared to the approach in [6], we propose to reduce the dimensionality of the feature vector using mean and standard deviation without compromising the classification accuracy. It is also important to point out the difference between sound event recognition as used in [6] and audio surveillance application as per the approach in [3]. In an audio surveillance application, a sound class has a number of different sound events. For example, shots fired from a rifle,

978-1-4799-4612-9/14/$31.00

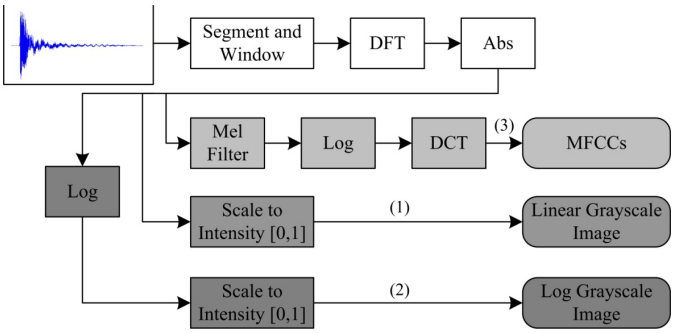

Figure 1. Steps in generating time-frequency images and determining MFCCs: (1) path for linear grayscale, (2) log grayscale, and (3) MFCCs

shotgun, and machine gun are sound events but in an audio surveillance application they are treated as a single sound class such as gunshots. In some cases, the signal properties of subclasses in a particular class are similar to the subclasses in other classes but different to subclasses in its own class as mentioned in [3]. This creates interclass similarity and intraclass diversity, increasing the complexity of the problem as a result.

The rest of this paper is organized as follows. Section II gives an overview of generating time-frequency images and feature extraction. Section III is on the experimentations we carried out and the corresponding results while conclusion and future recommendations are given in Section IV.

II. Feature Extraction

The procedure for generating the time-frequency images is described below together with feature extraction and computation of MFCCs which we use for comparing the results.

A. Grayscale Spectrogram

The generation of the linear and log grayscale spectrograms takes path 1 and 2 respectively, as shown in Fig. 1. Firstly, the Discrete Fourier Transform (DFT) is applied to the windowed signal as

$$X_t(k) = \sum_{n=0}^{N-1} x(n)w(n)e^{\frac{-2\pi ikn}{N}},\ k = 0,\ldots,N-1 \qquad (1)$$

where N is the window length, $x(n)$ is the time-domain signal, $X_t(k)$ is the k^{th} harmonic corresponding to the frequency $f(k) = kF_s/N$ for the t^{th} frame, F_s is the sampling frequency, and $w(n)$ is the window function.

The linear and log values are then obtained as

$$S_{Linear}(k,t) = \left|X_t(k)\right| \qquad (2)$$

$$S_{Log}(k,t) = \log\left|X_t(k)\right| \qquad (3)$$

These values are then normalized in the range [0,1] which gives the grayscale image intensity values. The normalization is given as

$$G(k,t) = \frac{S(k,t) - \min(S)}{\max(S) - \min(S)}. \qquad (4)$$

The linear and log grayscale images for a sound signal from one of the sound classes, *construction*, are given in Fig. 2.

Each image is then divided into blocks from which central moments are computed. The m^{th} central moment for any given block of image can be determined as

$$\mu_m = \frac{1}{K}\sum_{i=1}^{K}(G_i - \mu)^m \qquad (5)$$

where K is the sample size or the number of pixels in the block, G_i is the grayscale intensity value of the i^{th} sample in the block, and μ is the mean grayscale intensity value of the block.

B. MFCCs

The extraction of MFCCs follows path 3. A triangular mel filterbank is applied to the linear spectra and the energy in each filter is added. The discrete cosine transform (DCT) of the log power of these values are then computed from which the MFCCs are obtained.

III. Experimental Evaluation

A description of the database of sounds used in this work is given first followed by an overview of the noise conditions and the experimental setup. We then compare the classification accuracy using MFCCs and the time-frequency image-based features. We also present the confusion matrix to view the classification and misclassification of test samples from each class followed by the results for multi-conditional training.

A. Description of Sound Database

The sound database consists of 10 classes: *alarms*, *children voices*, *construction*, *dog barking*, *footsteps*, *glass breaking*, *gunshots*, *horn*, *machines*, and *phone rings*. The sound files are largely obtained from the Real World Computing Partnership (RWCP) Sound Scene database in Real Acoustic Environment [12] and the BBC Sound Effects library [13]. All signals in the database have 16-bit resolution and a sampling frequency of 44100 Hz. The choice of the sound classes is similar to most other audio surveillance applications, [3] in particular.

B. Noise Conditions

The performance of the different features and classification methods are investigated under three different noise environments taken from the NOISEX-92 database [14]: *speech babble*, *factory floor 1*, and *destroyer control room*. The signals are resampled at 44100 Hz and the overall performance is measured in clean conditions and at 20dB, 10dB, and 0dB signal-to-noise (SNR).

C. Experimental Setup

For all experiments, features were extracted from a Hamming window of 512 points (11.61 ms) with 50% overlap. We compare the results using SVMs and K-Nearest Neighbor

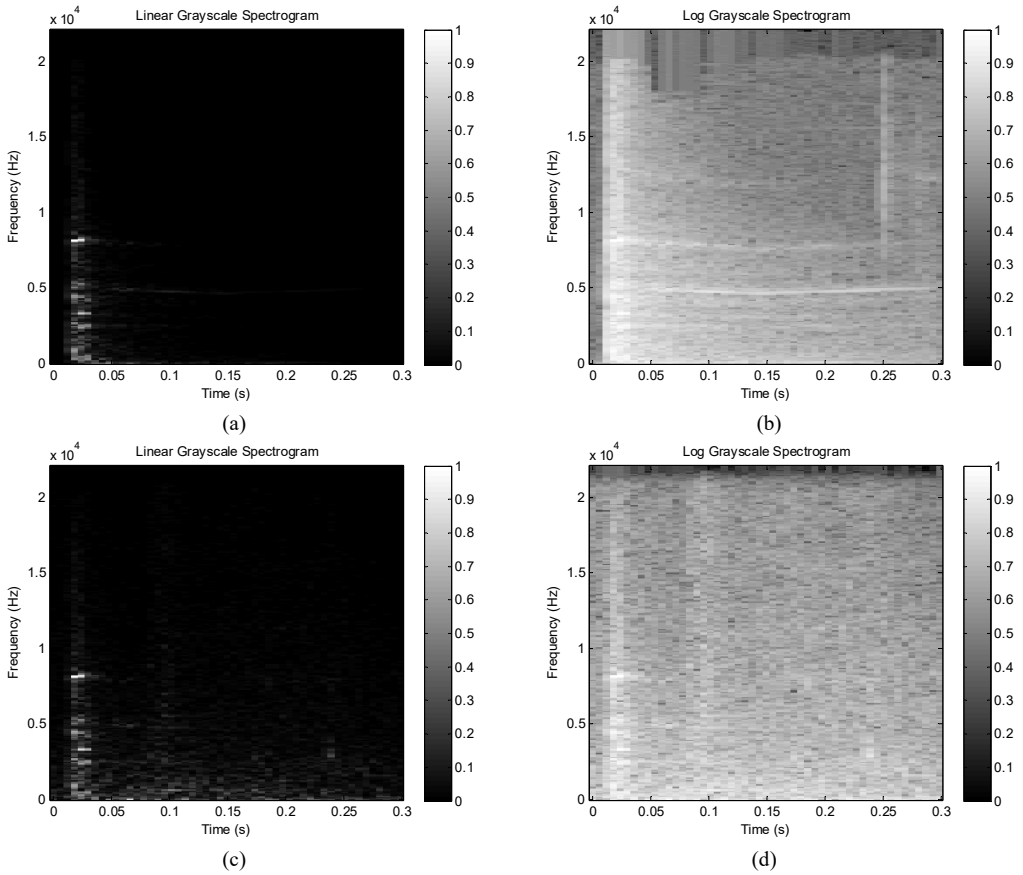

Figure 2. Grayscale images for a sound signal from *construction* sound class. (a) Linear grayscale image under clean conditions, (b) log grayscale image under clean conditions, (c) linear grayscale image at 0dB SNR with factory noise, and (d) log grayscale image at 0dB SNR with factory noise.

(KNN) classification. Four multiclass SVM classification techniques are used: one-against-all (OAA) [15, 16] where the classifier that has the highest output function assigns the class, one-against-one (OAO) using a max-wins voting strategy [17], decision directed acyclic graph (DDAG) [18], and adaptive directed acyclic graph (ADAG) [19]. All results reported are using a nonlinear SVM with a Gaussian radial basis function kernel as it was found to give the best results during preliminary experiments. The classifier parameters were tuned using cross validation.

The system is trained with two-third of the clean samples with all remaining data used for testing. Under multi-conditional training, two-third data from clean samples and at 0dB SNR are used for training while all remaining data is used for testing. For the MFCC method, the feature vector for each frame is 36-dimensional: 12 MFCCs with the 0^{th} component excluded, using a 23-filterbank system, plus deltas and accelerations. The overall size of the feature vector for a signal is $36 \times F$, where F is the number of frames in the sound signal, which is different in each case. After data normalization, the final feature vector is represented by concatenating the mean and standard deviation for each dimension. As such, the final feature vector is 72-dimensional.

For the time-frequency image feature method, the image is divided into 9×9 blocks and second and third central moments were computed in each block. We experimented with

TABLE I. CLASSIFICATION ACCURACY USING MFCCS - TRAINING USING CLEAN SAMPLES ONLY

Classification Method	MFCC				
	Clean	20dB	10dB	0dB	Average
OAA-SVM	98.43	90.81	69.03	41.56	**74.96**
OAO-SVM	98.16	90.52	65.65	36.48	72.70
DDAG-SVM	98.16	91.28	63.43	35.58	72.11
ADAG-SVM	98.16	91.89	65.12	37.12	73.08
KNN	96.59	87.17	57.63	31.73	68.28

$3 \times 3, 5 \times 5$, and 7×7 blocks as well but best results were obtained with 9×9 blocks. It was seen that the classification accuracy increased with an increase in the number of blocks but 9×9 was the maximum that could be experimented with due to limitations in the length of the sound signal and the size of the image as a result. As per the approach in [6], the size of the concatenated feature vector for the grayscale spectrograms is $9 \times 9 \times 2 = 162$ where two refers to the second and third central moments extracted from each of the 81 blocks.

We also experimented with concatenating the mean and standard deviation of the raw data along the row and column of the blocks to form the feature vector. Using this technique, the size of the feature vector for the grayscale spectrogram is $9 \times 4 \times 2 = 72$, where four represents the two pairs of mean and standard deviation for the second and third central moment. The classification accuracy using the second approach

TABLE II. CLASSIFICATION ACCURACY USING TIME-FREQUENCY IMAGE METHODS - TRAINING USING CLEAN SAMPLES ONLY

Classification Method	Linear Grayscale					Log Grayscale				
	Clean	20dB	10dB	0dB	Average	Clean	20dB	10dB	0dB	Average
OAA-SVM	91.34	97.35	94.63	55.82	**84.78**	95.28	68.42	45.67	30.42	**59.94**
OAO-SVM	91.86	90.52	85.30	49.66	79.34	95.54	66.99	44.53	29.10	59.04
DDAG-SVM	91.60	91.25	86.03	47.42	79.08	95.54	66.38	44.07	28.58	58.64
ADAG-SVM	91.34	90.93	87.23	51.76	80.31	95.54	67.40	44.88	29.75	59.39

TABLE III. CONFUSION MATRIX UNDER CLEAN CONDITIONS USING LINEAR GRAYSCALE IMAGE METHOD WITH OAA-SVM CLASSIFICATION

	Alarms	Children Voices	Construction	Dog barking	Footsteps	Glass breaking	Gunshots	Horn	Machines	Phone rings
Alarms	96.67	3.33	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Children voices	3.33	80.00	0.00	1.67	3.33	3.33	0.00	5.00	1.67	1.67
Construction	0.00	3.33	96.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Dog barking	7.14	7.14	0.00	85.71	0.00	0.00	0.00	0.00	0.00	0.00
Footsteps	0.00	7.02	0.00	0.00	92.98	0.00	0.00	0.00	0.00	0.00
Glass breaking	0.00	5.00	0.00	0.00	0.00	95.00	0.00	0.00	0.00	0.00
Gunshots	0.00	7.14	0.00	0.00	0.00	0.00	92.86	0.00	0.00	0.00
Horn	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00	0.00
Machines	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	100.00	0.00
Phone Rings	0.00	15.22	0.00	0.00	0.00	0.00	0.00	0.00	0.00	84.78
Overall Classification Accuracy = 91.34										

TABLE IV. CONFUSION MATRIX AT 0dB SNR USING LINEAR GRAYSCALE IMAGE METHOD WITH OAA-SVM CLASSIFICATION

	Alarms	Children Voices	Construction	Dog barking	Footsteps	Glass breaking	Gunshots	Horn	Machines	Phone rings
Alarms	39.26	41.30	0.00	0.74	0.00	0.93	0.00	5.00	2.96	9.81
Children voices	0.93	79.26	0.74	0.93	1.48	1.11	0.93	6.67	5.74	2.22
Construction	1.48	22.59	71.11	0.00	0.00	0.00	2.59	0.37	0.37	1.48
Dog barking	2.38	47.22	0.00	32.94	0.79	0.00	0.00	9.13	7.54	0.00
Footsteps	0.00	21.44	0.19	0.00	71.54	0.19	0.00	5.46	0.00	1.17
Glass breaking	0.00	5.00	4.44	0.00	0.00	89.44	0.00	1.11	0.00	0.00
Gunshots	0.00	49.60	0.40	4.76	0.00	0.40	24.60	2.78	0.00	17.46
Horn	0.00	33.33	0.00	0.00	1.01	0.00	0.00	62.63	2.02	1.01
Machines	0.00	29.63	0.00	1.48	0.00	4.81	0.00	3.33	47.78	12.96
Phone Rings	0.00	17.87	0.00	0.00	0.48	0.00	0.00	13.29	30.68	37.68
Overall Classification Accuracy = 55.82										

was slightly better and with the added advantage of reduced dimensionality for the feature vector, we present results using this method only.

D. Results

The classification accuracy with MFCCs are given in Table I. The minimum classification accuracy in clean conditions is 98.16% for the SVM methods and is 96.59% for KNN. However, the classification accuracy reduces greatly with the addition of noise, especially at 10dB and 0dB SNR with the highest classification accuracy at 69.03% and 41.56%, respectively. In general, it was seen that the multiclass SVM classification methods produced better results than KNN so from here on we present the results using the multiclass SVM classification methods only.

The classification accuracy using the time-frequency image feature approach is given in Table II. While the log grayscale method gives better classification accuracy under clean conditions, the linear grayscale method performs much better under noisy conditions. The same comparison also applies with the results obtained using MFCCs. As such, in terms of overall performance, the linear grayscale approach outperforms the log grayscale and MFCC methods. Also, in all the cases, OAA-SVM gives a better overall performance than the other three methods. The results are especially better under noisy conditions.

The log grayscale approach can be expected to perform better in clean conditions since taking log power reveals the details in the low power frequencies unlike the linear grayscale approach where only the dominant power frequencies are shown. This can be visualized in the linear grayscale and log grayscale images in Figure 2(a) and (b), respectively. However, the performance of the two representations changes with the addition of noise. The noise is more diffuse than the sound signal and its power affects most of the frequencies in the log grayscale image as shown in Figure 2(d). For the linear representation, the strong peaks of the sound are larger than the noise and remain largely unaffected with the addition of noise as can be seen in Figure 2(c).

TABLE V. CLASSIFICATION ACCURACY USING TIME-FREQUENCY IMAGE METHODS - MULTI-CONDITIONAL TRAINING

Classification Method	Linear Grayscale					Log Grayscale				
	Clean	20dB	10dB	0dB	Average	Clean	20dB	10dB	0dB	Average
OAA-SVM	91.34	96.59	96.33	91.43	**93.92**	92.39	80.52	84.69	85.83	85.86
OAO-SVM	91.34	96.85	95.36	90.99	93.64	92.39	81.69	85.16	84.60	**85.96**
DDAG-SVM	91.08	96.82	95.48	91.08	93.61	92.91	81.22	84.31	84.16	85.65
ADAG-SVM	90.81	96.73	95.39	91.60	93.64	92.39	81.16	84.69	84.34	85.64

Table III and IV present the confusion matrices for the OAA-SVM method under clean conditions and at 0dB SNR, respectively, for the linear grayscale image method. The confusion matrix allows the observation of the degree of confusion between the different classes which gives a better understanding of the classification performance when compared to the overall classification accuracy results given in Table II. The rows of the confusion matrix denote the sound classes that we want to classify and the columns denote the classified results. The values are given in percentage as *number of correctly (or incorrectly) classified samples* divided by *number of test samples in the class*.

As an example, for the confusion matrix under clean conditions given in Table III, for the test samples from *alarms*, 96.67% were correctly classified as *alarms* while the remaining 3.33% were incorrectly classified as *children voices*. It can be said that test samples from *children voices* were more often misclassified than the other sound classes. Apart from *dog barking*, *children voices* is the only other class which has more than one confusion and it has confusion with all the other classes except *construction* and *gunshots*. We can also say that there is one sided confusion between *construction* and *children voices* where test samples from *construction* were misclassified as *children voices* but not vice-versa. There is also one sided confusion between *gunshots* and *children voices*. *Horn* and *machines* are the best performing classes with no misclassifications.

While *children voices* has the lowest classification accuracy under clean conditions at 80.00%, interestingly, it has one of the highest classification accuracies at 0dB SNR at 79.26%, as per the results in Table IV. It is only behind *glass breaking*, which at 89.44%, has the highest classification accuracy. *Children voices* also has the smallest reduction in the classification accuracy although it now has misclassifications with all the other classes. In addition, all the other classes have misclassifications in *children voices* and *horn*. *Alarms*, *dog barking*, *gunshots*, *machines*, and *phone rings* are the worst performing classes with a classification accuracy of less than 50%.

In addition, we present the classification accuracy with multi-conditional training in Table V. The linear grayscale approach once again gives the best overall classification accuracy which is with the OAA-SVM classification method. The most improved result using this approach is at 0dB SNR, from 55.82% when trained with clean samples only to 91.43% with multi-conditional training. However, the disadvantage of using multi-conditional training is that the number of training samples has increased fourfold which increases the training time

IV. CONCLUSION

The linear grayscale approach was found to be more robust than the log grayscale approach under noisy conditions. While the proposed method outperforms the conventional approach of using MFCCs as features, there is still room for improvement in the classification accuracy especially at low SNRs. In this work, we considered central moments as features but other distribution statistics could be explored in future work.

REFERENCES

[1] G. Guo and S. Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Transactions on Neural Networks,* vol. 14, no. 1, pp. 209-215, 2003.

[2] S. Chu, S. Narayanan, and C. C. J. Kuo, "Environmental sound recognition with time-frequency audio features," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 17, no. 6, pp. 1142-1158, 2009.

[3] 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.

[4] X. Zhang and Y. Li, "Environmental sound recognition using double-level energy detection," *Journal of Signal and Information Processing,* vol. 4, no. 3B, pp. 19-24, 2013.

[5] E. Alexandre, L. Cuadra, M. Rosa, and F. Lopez-Ferreras, "Feature selection for sound classification in hearing aids through restricted search driven by genetic algorithms," *IEEE Transactions on Audio, Speech, and Language Processing,* vol. 15, no. 8, pp. 2249-2256, 2007.

[6] J. Dennis, H. D. Tran, and H. Li, "Spectrogram image feature for sound event classification in mismatched conditions," *IEEE Signal Processing Letters,* vol. 18, no. 2, pp. 130-133, 2011.

[7] K. Abe, H. Sakaue, T. Okuno, and K. Terada, "Sound classification for hearing aids using time-frequency images," in *IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PacRim)*, 2011, pp. 719-724.

[8] Y. M. G. Costa, L. S. Oliveira, A. L. Koericb, and F. Gouyon, "Music genre recognition using spectrograms," in *18th International Conference on Systems, Signals and Image Processing (IWSSIP)*, 2011, pp. 1-4.

[9] C. N. Silla Jr., A. L. Koerich, and C. A. A. Kaestner, "The Latin music database," in *9th International Conference on Music Information Retrieval (ISMIR)*, Philadelphia, PA, USA, 2008, pp. 451-456.

[10] R. M. Haralick, "Statistical and structural approaches to texture," *Proceedings of the IEEE,* vol. 67, no. 5, pp. 786-804, 1979.

[11] M. Lopes, F. Gouyon, A. L. Koerich, and L. E. S. Oliveira, "Selection of training instances for music genre classification," in *20th International Conference on Pattern Recognition (ICPR)*, 2010, pp. 4569-4572.

[12] S. Nakamura, K. Hiyane, F. Asano, T. Nishiura, and T. Yamada, "Acoustical sound database in real environments for sound scene understanding and hands-free speech recognition," in *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC 2000)*, Athens, Greece, 2000, pp. 965–968.

[13] *BBC Sound Effects Library*. Available: http://www.leonardosoft.com

[14] A. Varga and H. J. M. Steeneken, "Assessment for automatic speech recognition: II. NOISEX-92: A database and an experiment to study the effect of additive noise on speech recognition systems," *Speech Communication,* vol. 12, no. 3, pp. 247-251, July 1993.

[15] V. N. Vapnik, *Statistical learning theory*. New York: Wiley, 1998.

[16] L. Bottou, C. Cortes, J. S. Denker, H. Drucker, I. Guyon, L. D. Jackel*, et al.*, "Comparison of classifier methods: a case study in handwritten digit recognition," in *Proceedings of the 12th IAPR International Conference on Pattern Recognition, Vol. 2 - Conference B: Computer Vision & Image Processing*, 1994, pp. 77-82.

[17] U. H. G. Kreßel, "Pairwise classification and support vector machines," in *Advances in Kernel Methods - Support Vector Learning*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA: MIT Press, 1999, pp. 255-268.

[18] J. C. Platt, N. Cristianini, and J. Shawe-Taylor, "Large margin DAGs for multiclass classification," in *Advances in Neural Information Processing Systems 12 (NIPS-99)*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds. Cambridge MA: MIT Press, 2000, pp. 547-553.

[19] B. Kijsirikul, N. Ussivakul, and S. Meknavin, "Adaptive directed acyclic graphs for multiclass classification," in *PRICAI 2002: Trends in Artificial Intelligence*. vol. 2417, M. Ishizuka and A. Sattar, Eds. Berlin Heidelberg: Springer, 2002, pp. 158-168.