REVIEW OF ICA AND HOS METHODS FOR RETRIEVAL OF NATURAL SOUNDS AND SOUND EFFECTS.

Shlomo Dubnov

Department of Communication Engineering
Ben-Gurion University
Be’er-Sheva,Israel
dubnov@bgumail.bgu.ac.il

Adiel Ben-Shalom

School of Computer Science
Hebrew University
Jerusalem, Israel.
chopin@cs.huji.ac.il

ABSTRACT

Search by similarity in sound effects is needed for musical authoring and search by content (MPEG-7) applications. Sounds such as outdoor ambience, machine noises, speech or musical excerpts as well as many other man made sound effects (so called ”Folley sounds”) are complex signals that have a well perceived acoustic characteristic of some random nature. In many cases, these signals can not be sufficiently represented based on second order statistics only and require higher order statistics for their characterization. Several methods for statistical modeling of such sounds were proposed in the literature: non-gausian linear and non-linear source-filter models using HOS, optimal basis / sparse geometrical representations using ICA and methods that combine ICA-based features with temporal modeling (HMM). In this paper we review several such approaches and evaluate them in the context of multimedia sound retrieval.

1. INTRODUCTION

The need for finding similarity between sounds appears in diverse applications, such as acoustic monitoring, medical diagnosis, analysis of animal vocalizations and multimedia content description [1],[2],[3]. Audio content in multimedia is broadly divided, according to tradition of post-production practice, into categories of speech, music and Foley sounds (sound effects). Huge body of literature exists on the topic of speech analysis, the main problems being speaker identification, language recognition and speech transcription [4]. These methods usually rely on a set of well established features that are appropriate for speech modeling, such as cepstral coefficients and cepstral derivatives. The modeling of a single source (such as a specific speaker) is usually achieved using parametric modeling of the distribution of these features, such as using Gaussian Mixtures (GMM) [5].

Music identification task concerns finding similarity between musical pieces based on mutliple features, such as pitch [6], beat [7] and multiple features derived from timbre (sound color) analysis [8]. The many applications of music similarity include genre classification, query by example and music thumbnailing. The problem of discrimination between speech, music and sound effects, are considered in [9],[10]. The authors used Gaussian Mixture Model (GMM) and Hidden Markov Model (HMM) to estimate a statistical model of the signal spectrum distribution.

Sounds that have received the least attention among the three categories are the sound effects. One of the reasons for that might be the lack of a clear definition of the sound effects category, which might include diverse sounds ranging from simple colored noise to most complex acoustic scenarios. This situation makes it difficult to design a-priori, knowledge based features that are derived from some physical motivation about the nature of the sound and requires to consider more generic statistical data-based modeling approaches. Moreover, this situation does not allow to define one set of acoustic features that would fit all cases. Additional difficulty is in the lack of clear signal-related taxonomy for performing the classification or construction of statistical models. Sound categories are usually derived based on the context or a broad definition of the sound source rather then its signal properties. For instance the category of car sounds may include different engine noises, ignition sounds, breaks, door slam or car crash, which have no evident signal similarity.

In this paper we will review several method for classification of sound effects that employ higher order statistics or spectra (HOS) [11] and independent components analysis (ICA) [12], for finding similarity between the sounds. We broadly distinguish between two types of methods. The first assumes non-gausian linear and non-linear source-filter models for the signal and use HOS for measuring similarity between the signal. A second type of methods searches for optimal basis / sparse geometrical representations of the amplitude spectrum using ICA for each sound class. This approach represents the sound as a combination of a small set of independent spectral components. Classification is achieved by evaluating the likelihood of a new sound among a set of trained models. Two approaches for likelihood estimation will be described in the paper. First, the likelihood is maximized by searching for a model that gives minimal mutual information among ICA coefficients of ICA models of the different classes. These results are compared to another ICA-based method that employs temporal modeling of the coefficient distribution by means Hidden Markov Model (HMM). The trained HMM's for each class are then used for classification of ICA coefficients.

The applications to be considered are search by similarity / query from example and sound classification given a taxonomy or a labeled training database that is partitioned into similarity classes.

2. HIGHER ORDER STATISTICAL FEATURES

When signals have Gaussian density distributions, we can describe them thoroughly with second order measures like the autocorrelation function or the spectrum. In the case of noisy signals such as engine noises of sound effects, the second order statistics are not sufficient for a good signal characterization and matching. Assuming a source-filter model, non-Gaussian statistics in the signal may appear if

- A linear system is excited by a non-Gaussian source signal.
- The system itself is non-Linear. I n such a case couplings may occur between signal components at different frequencies.

Using, for instance, the bicoherence statistic, the first case is characterized by constant non-zero bicoherence, while the second case has varying shape of bicoherence, with peaks corresponding to coupled bi-frequencies.

Using HOS statistics as features for signal classification, several authors suggested to match image textures [13] and audio textures [14] by matching higher order cumulants or polyspectra. In this approach, the higher order statistics are measured directly from the signal and are used as a feature for matching or classification instead of spectral features. Direct matching using polyspectral features also captures differences due to non-linear characteristics of the signal.

2.1. Higher Order Spectral Distance Measure

One can see that the bispectral amplitudes of the signals differ significantly. In order to perform signal recognition one needs a new measure for capturing the similarity/differences between the HOS properties of the various signals. We develop a higher order statistical distance measure, which comes in time-domain or spectral flavors: Lets us denote by $\hat{c}_y^k(\tau_1, \tau_2, ..., \tau_{k-1})$ a higher order statistics vector that contains k-th order cumulant lags of signal y. The probability density of the estimates is known to be approximately normal around the true cumulant value. This allows one to compute the conditional probability of statistics of y given the cumulants of another signal x

$$\begin{aligned} p(\hat{c}_y^k | c_x^k) &= (2\pi)^{-L/2} |\frac{1}{N}\Sigma_x|^{-1/2} \qquad (1) \\ & \quad \cdot exp\{-\frac{1}{2}[\hat{c}_y^k - c_x^k]^T [\frac{1}{N}\Sigma_x|^{-1} [\hat{c}_y^k - c_x^k] \end{aligned}$$

where N is the length of the signal and L the maximal lag for which the higher order statistics are estimated. This allows us to write the cross-entropy between the two signals $D(c_y^k, c_x^k) =< log\frac{p_y}{p_x} >_{p_x}$, where we use the estimates in place of the true statistics. In frequency domain it can be shown that an equivalent objective becomes

$$\begin{aligned} d_{HOS}(C_y, C_x) = \frac{\gamma_k}{2} \sum_{(\omega_1,...,\omega_{k-1}) \in \Omega_k} & |C_{\tilde{y}}^k(\omega_1, ..., \omega_{k-1}) \qquad (2) \\ & - C_{\tilde{x}}^k(\omega_1, ..., \omega_{k-1})|^2 \end{aligned}$$

$$\begin{aligned} & C_{\tilde{x}}(\omega_1, \omega_2, ..., \omega_{k-1}) \qquad (3) \\ & \approx \frac{C_x(\omega_1, \omega_2, ..., \omega_{k-1})}{\sqrt{S_x(\omega_1) S_x(\omega_2) ... S_x(\omega_1 + ... + \omega_{k-1})}} \end{aligned}$$

with C_x^k being k-th order polyspectra of signal x, estimated over non-redundant poly-frequency region Ω_k and γ_k a constant that depends on the windowing method. Note that $y(t)$ is decorrelated by a filter that matches the spectrum of $x(t)$, i.e. matches S_x. This higher order statistical distance measure is a Kullback-Leibler (KL) divergence between probability distributions of the higher order statistics (HOS) vectors of the signals in channels y and x[1].

3. EMBEDDED REPRESENTATION AND INDEPENDENT COMPONENTS

Another approach considers a geometrical embedding of the signal in a transformed space. Assuming that the signal is represented as a sequence of independent linear combinations of n dimensional basis vectors, decomposition of the signal into "channels" or "expansion coefficients" of the transformed space is performed. This approach is common to many signal compression schemes such as transform or sub-band coding. Given a multi-variate distribution of vectors $\hat{x} = (x_1, x_2, ..., x_n)$, we want to find a matrix W and vector $\hat{s} = (s_1, s_2, ..., s_n)$ so that the components of the vector are "as independent as possible". In other words, it is assumed that exists a multivariate process with independent

[1] It is important to note that this not KL distance between the signals s_y, s_x themselves but rather between their HOS statistics c_y, c_x.

components $\hat{s}$ and a matrix $A = W^{-1}$, so that $\hat{x} = A\hat{s}$. The representation can be obtained by applying ICA to one-dimensional time signal that is embedded in n-dimensional blocks.

In order to obtain an initial sparse representation we transform the signal vectors into a frequency domain. A major departure from the signal based methods is that we are considering the magnitude spectrum, rather then the complex spectrum. This second step discards redundancies in the signal due to possible time shifts. Same invariance can be obtained by using a filter-bank instead of a simple FFT and taking one of the two quadrature components of every filter in the filter-bank. We shall denote the magnitude spectrum vector by X. For each sound we obtain a sequence of such vectors over time[2].

A third step that is employed before ICA modeling is a data reduction step. We do it by reducing the dimension of the row space of $\mathbf{X^T}$ by using the singular value decomposition (SVD) method. $\mathbf{X^T}$ can be decomposed to :

$$\mathbf{X^T} = \mathbf{USV^T} \tag{4}$$

where U is an $m * m$ matrix and V is an $n * n$ matrix and S is a diagonal matrix which contains the singular values of $\mathbf{X^T}$. In our scheme, m stands for the number of time observations and $n = 128$ is the number of FFT bins. To reduce the dimension of the row space of $\mathbf{X^T}$ to a lower dimension r, we project $\mathbf{X^T}$ on the first r column vectors of $\mathbf{V}$

$$\mathbf{Y^T} = \mathbf{X^T V_r} \tag{5}$$

where $\mathbf{V_r}$ is a matrix which contains the first r column vectors from $\mathbf{V}$. A reduced dimension $r = 10$ was chosen in our experiments.

3.1. ICA features for Sound Representation

Bell & Sejnowski in [15] apply ICA for feature extraction for natural audio signals. They seek for basis sound waveforms, which can be thought of basis functions for natural sounds. Short sound segments of this sound clip (about 20ms each) were taken and organized in the columns of the observation matrix. Then ICA analysis was done on the observation matrix to extract the basis functions and the statistically independent weights on these basis functions, which together comprise the sound clip segments. The superior behavior of ICA over second order methods such as PCA can be explained by the fact that second order methods reflect only the amplitude spectrum of a signal and ignore the phase spectrum. The local features of a signal are reflected in the phase spectrum, thus second order statistics methods cannot reveal this information. On the other hand the ICA algorithm depends not only on the second order statistics of the signal but also on higher order statistics. This reflects not only the spectrum of the signal but also the phase information. Thus, ICA method extracts local features in the signal, which are not detected by second order tools.

Casey in [16] uses different architecture to extract features from sound files. Unlike the signal decomposition used by Bell & Sejnowski, the decomposition of the sound according to Casey should not be done on the signal waveform. The signal is first transformed to the spectro-temporal domain using Short Time Fourier transform (STFT). This is justified by the fact that most of the salient information in audio signals exists in the short-time spectro-temporal domain. The ICA decomposition is done only on the magnitude part of the frequency domain representation and the phase information is omitted.

3.2. Naive Bayesian Classifiers

Once the probabilities for the classes and the conditional probabilities for a given feature and a given class are estimated, this information can be used to classify each new instance. Usually the estimation is based on the frequency of occurrence of the features and classes over the training data, simply by counting the frequency of various data combinations within the training examples. This type of classifier is called nave because it assumes independency of the features. This technique has been used for classification of 30 short sounds of oboe and sax using 18 Mel-Cepstrum Coefficients, with an accuracy rate of 85% [17]. After clustering the feature vectors with a K-means algorithm, a Gaussian mixture model was used to estimate a parametric representation of the probabilities for a Bayessian classifier. A more complex approach that combines temporal (transition) statistics with ICA feature extraction is presented in [18]. The algorithm tries to classify between N classes of sounds. For each class a separate HMM is trained on samples from that class. After the training step, the algorithm is given a sound sample which wasn't in the training set and it estimates which HMM model is the most likely model which generated this sound sample.

Next we consider a new classifier that departs from the independence assumption and uses the relation between feature mutual-information and the data likelihood in ICA context. This is compared to HMM based classifier.

4. CLASSIFICATION BY LEAST CROSS-FEATURE MUTUAL INFORMATION OF ICA FEATURES

Assuming that the sound can be successfully modeled as a combination of independent spectral vectors, a know simple relation exists between the likelihood of the data and the

[2] This is equivalent to a spectrogram or magnitude of short-time Fourier transform (STFT) representation of the signal.

mutual information between the independent components. Let us denote by $P(x|W)$ the hypothesized distribution of the data given a model W. We assume that the data vector x is modeled by a linear combination with weights s of basis vectors A_i. Written in matrix form, $x = As$ (and the inverse relation will be denoted as $s = Wx$). In the assumed model, the mean log-likelihood of the data can be approximately written as

$$logL \approx -\sum H(s_i) - log|det(A))| \; . \tag{6}$$

The relation between the "true" entropy of the signal and the coefficents s is

$$H(x) = H(s) + log|det(A)| \; . \tag{7}$$

Combining the two equations we get

$$logL + H(x) \approx -\sum H(s_i) + H(s) \tag{8}$$

which allows us to write the log-likelihhod as a function of the mutual information between the ICA coefficients

$$logL = -I(s) - H(x) \; . \tag{9}$$

Since $H(x)$ is independent of the model A, we can compare the likelihoods between different models by comparing $I(s)$. The most likely model will be the one with least mutual information between the coefficients $I(s)$.

Since in practice we can not measure mutual information between multiple variables (this requires construction of a high-dimensional histograms which is impractical with limited data) we used a sum of pairwise mutual informations $\sum I(s_i, s_j)$ as our likelihood measure. This provides higher limit to the true $I(s)$.

Our training and classification procedure can be summarized as follows:

- For each set of sounds in the training set, construct a short time magnitude spectrum matrix.
- After doing a dimension reduction of the spectrum matrix to a feasible size (in our case 10 vectors), subject the data to an ICA analysis. This resulting whithening matrix W (the inverse of the ICA basis) will serve as the model for the class.
- Given a test example, get the independent components for each class in consideration. This is achieved by simply multiplication of the test data with the saved matrices W in each class.
- Test the resulting coefficients s for independence by calculation of all pairs of $I(s_i, s_j)$ for each class. Decide that the class with highest likelihood is the one with the least mutual information between the coefficients

5. EXPERIMENTAL RESULTS

Two sets of experiments were conducted to test the HOS and ICA retrieval procedures: Query by Example and Sound Classification into a predetermined taxonomy of sound classes.

5.1. Query by similarity

Query by similarity was performed using HOS distance measure. Our implementation of the polyspectral matching is based on the equivalence between polyspectrum of decorrelated signal $\tilde{x}$ and a "spectrally normalized" version of polyspectra [3] of the original signal x (Eq. 3).

In order to avoid the need for long averaging to get reliable estimate of polyspectral features we used a k-th power matched filter procedure, briefly described below, that provides an upper limit to the polyspectral matching function. It is shown in [14] that

$$\begin{aligned} \tilde{d}(y,x) \;&=\; \int\int |C_{\tilde{x}}(\omega_1, \omega_2, ..., \omega_{k-1}) \\ &\qquad - C_{\tilde{y}}(\omega_1, \omega_2, ..., \omega_{k-1})|^2 \\ &\le\; \int (\tilde{x}(t) \otimes \tilde{x}(-t))^k dt - 2\int (\tilde{x}(t) \otimes \tilde{y}(-t))^k dt \\ &+\; \int (\tilde{y}(t) \otimes \tilde{y}(-t))^k dt \end{aligned} \tag{10}$$

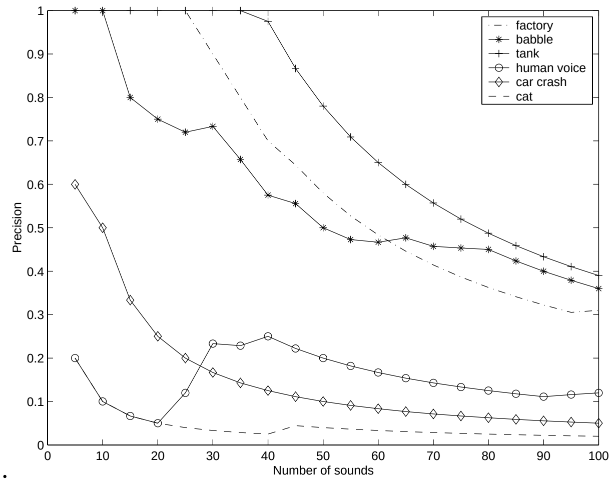
The database contained ≈ 500 files from different sound effects (SFX) classes, several machine noises [19], and other collections of sounds (e.g. phone, car crash, glass break). The decorrelation was done using LPC spectral estimation. We computed LPC whitening parameters (over 2 sec.) for each class using the first file in the class and saved it in different directory. When a new signal x is to be queried, we go over all the signals y in the SFX database, for y we taking the appropriate class LPC parameters (according to the SFX name) and decorrelating with it both the sound to be matched x and the queried sound y. We then measure $d(x,y)$ using the k-th power matched filter equation. The matching is done over each segment in the tested file to all segments in the matched file.

The results of query by example are summarized in the following graph which depicts the percentage of relevant sound files that were returned out of all sounds returned by the query. It can be seen that babble, factory and tank which are all noises got the best retrieval rate, while retrieval precision of other sounds such as cat, human voice and car crash was not so good.

5.2. Classification experiments

Classification experiments were done using ICA features and the two methods for likelihood evaluation: the Least

[3] For the case of $k = 3$, this decorrelated bispectrum measure is known as *bicoherence*.



..

Fig. 1. Query by Similarity results. The precision graph depicts the percentage of relevant sounds out of all sounds retrieved from the query.

Mutual Information (LMI) and HMM Training methods. For every sound in each class we performed spectrogram analysis with fft of 256. We computed ICA with data reduction to 10 components to get the appropriate W_i matrix for each class (i being the index of the class). In order to classify a new file we perform spectrogram analysis and then calculate $Y_i = W_i X$ to get the estimated ICA coefficients for candidate class i. In the LMI method we measure mutual information between the columns of Y_i in each class (10 vectors in each class) and choose the class with minimal mutual information. The method for estimating pairwise mutual information is based on non-parametric histrogram estimation of the mutual pairwise distribution for each pair of coefficients (row vectors) of Y_i [20]. In HMM method we train a HMM for each class using standard HMM training methods. Then the sequence of Y_i vectors (column vectors) are given to the HMM in order to obtain the global likelihood.

The database in the experiment contained 10 classes as is showm in the table, with ≈ 300 sounds. The instruments sounds were recorded from commercial CD's and each contained approximately 10 seconds of playing. We used 70% of the files for training and 30% for testing. All files are at 16Khz sampling rate.

The classification results are shown in the following tables. The LMI method yields good results for sounds such as cheering which can be characterized as stationary sounds. We also got good results when we used LMI method for classifying the SPIB database (results are not shown here). These sounds can be characterized by their texture which remains the same along the sound. For non-stationary signals such as male and female voice the addition of a dynamic model yields better results. An interesting point is that the HMM classifier failed to classify all guitar sound samples and it classified them to cello class. This might be because the guitar and cello sound samples that we chose resembled each other throughout the sound period. The LMI method was able to achieve better classification results on these classes.

	HMM			
	#Files	#Success	#Fail	Percentage
Violin	8	8	0	100%
Cello	11	11	0	100%
Guitar	7	0	7	0%
Piano	12	9	3	75%
Total	38	28	10	73%
Male	9	9	0	100%
Female	9	9	0	100%
Total	18	18	0	100%
Cheering	7	6	1	85%
Cat	2	0	2	0%
Glass Break	4	4	0	100%
Laughter	2	0	2	0%
Total	15	10	5	66%

Table 1. Classification results for ICA and dynamic model (HMM)

	Mutual Information			
	#Files	#Success	#Fail	Percentage
Violin	8	7	1	87%
Cello	11	9	2	81%
Guitar	7	7	0	100%
Piano	12	8	4	66%
Total	38	31	6	81%
Male	9	7	2	77%
Female	9	7	2	77%
Total	18	14	3	77%
Cheering	7	6	1	85%
Cat	2	1	1	50%
Glass Break	4	2	2	50%
Laughter	2	1	1	50%
Total	15	10	5	66%

Table 2. Classification results for ICA and Mutual Information Classifier

6. CONCLUSIONS AND FUTURE RESEARCH

The main differences between the HOS and ICA approaches are the complexity and time span of their underlying statistical model: in ICA the localization property favors decomposition into independent events that could be differently recombined during the course of the sound , while source-filter models try to capture the complete statistics in one model, thus requiring stronger stationarity assumption. Our experiments indicate that a very good matching can be obtained for the case of stationary signals. In such a case HOS distance can be applied for query and classification tasks. In reality, most of the SFX's used in multimedia applications are non-stationary. When the signal is composed of several sound events, this situation offers an advantage for ICA methods that separately capture different component statistics. The importance of this separation is supported by our findings that classification can be based on testing the coefficients independence assumption (LMI classifier) rather then explicit modeling and matching between the statistics of the ICA coefficients of the test sound and the target classes, as is done in the HMM-based classifier.

It is importance to note that the LMI method, as developed in this paper, assumes that an exact model W is available for the target sound classes. This allows to calculate the data likelihood from the Mutual Information function among coefficients only. When an error between the model and the true distribution is included, and additional modeling error term should be included. In our case we believe that at least part of the model errors was avoided due to the PCA data reduction step (this would be true if the model errors could be considered as some sort of additive noise). Better model robustness and error resilience can be considered, using adaptive PCA thresholding or adaptive coefficient shrinkage methods. Additional issue to be considered are faster classification and matching algorithm, including efficient algorithms for Mutual Information and HOS distance measure calculation, fast search methdos and hierarchical representation for the sound class taxonomy.

7. ACKNOWLEDGMENTS

This research was partially funded by the CUIDADO European project.

8. REFERENCES

[1] D. Keislar E. Wold T. Blum and J. Wheaton, "Contentbased classification, search and retrieval of audio," *IEEE Multimedia*, vol. 3, no. 2, pp. 27–36, 1996.

[2] P. Herrera, X. Serra, and G. Peeters, "Audio descriptors and descriptor schemes in the context of mpeg-7," 1999.

[3] Sikora T. (Editors) Manjunath B.S., Salembier P., *Introduction to MPEG-7: Multimedia Content Description Interface*, Wiley, 2002.

[4] L.R. Rabiner and B.H Juang, *Fundamentals of Speech Recognition*, Prentice Hall, 1993.

[5] D.A. Reynolds and R.C. Rose, "Robust text independent speaker identification using gaussian mixture speaker model," *IEEE Trans. on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[6] A. Ermolinski G. Tzanetakis and P. Cook, "Pitch histograms om audio and symbolic music information retrieval," in *Proceedings of the International Conference on Music Information Retrieval, ISMIR 2002, Paris*, 2002.

[7] A. Klapuri J. Paulus, "Measuring the similarity of rhytmic patterns," in *Proceedings of the International Conference on Music Information Retrieval, ISMIR 2002, Paris*, 2002.

[8] K. Jensen and J. Arnspang, "Binary tree classification of musical instruments," in *Proceedings of the ICMC, Beijing, China*, 1999.

[9] T. Zhang and C. Kuo, "Hierarchical system for contentbased audio classification and retrieval," .

[10] Lie Lu, Hao Jiang, and HongJiang Zhang, "A robust audio classification and segmentation method," in *ACM Multimedia*, 2001, pp. 203–211.

[11] A. Petropulu C.L. Nikias, *Higher-Order Spectral Analysis: A Non-Linear Signal Processing Framework*, Prentice Hall, 1993.

[12] Hyvarinen, "Survey on independent component analysis," *Neural Computing Surveys*, , no. 2, pp. 94–128, 1999.

[13] M.K.Tsatsanis and G.B.Giannakis, "Object and texture classification using higher order statistics," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 14, no. 7, 1992.

[14] S. Dubnov and N. Tishby, "Analysis of sound textures in musical and machine sounds by means of higher order statistical features," in *Proceeding of International Conference on Acoustics Speech and Signal Processing, Munich*, 1997.

[15] A. J. Bell and T. J. Sejnowski, "Learning the higher-order structure of a natural sound," *Network: Computation in Neural Systems*, 1996.

[16] M.A. Casey, *Auditory Group Theory: with Applications to Statistical Basis Methods for Structured Audio*, Ph.D. thesis, MIT Media Lab, 1998.

[17] Brown J.C., "Computer identification of musical instruments using pattern recognition with cepstral coefficients as features," *J. Acoust. Soc. Am. 105, 1933-1941*, 1999.

[18] M. Casey, *Sound Classification and Similarity Tools, in B.S. Manjunath, P. Salembier and T. Sikora, (Eds), Introduction to MPEG-7: Multimedia Content Description Language,*, J. Wiley, 2001.

[19] SPIB database:, "http://spib.rice.edu/spib/select_noise.html," .

[20] R. Moddemeijer, "On estimation of entropy and mutual information of continuous distributions," *Signal Processing*, vol. 16, no. 3, pp. 233–246, 1989.