A Robust Environmental Sound Recognition System using Frequency Domain Features

T. Sivaprakasam
Assistant Professor
Dept. of Computer Science & Engg.
Annamalai University, India

P. Dhanalakshmi
Associate Professor
Dept. of Computer Science & Engg.
Annamalai University, India

ABSTRACT
In ubiquitous environments, analysis and classification of sound plays a critical role in various acoustic-based recognition systems. This work aims to contribute towards building an automatic sound recognition system that can understand the surrounding environment by the audio information. In this paper, an acoustic signal based context awareness system is proposed for detecting sound events in five different real-world environment.This approach is based on Back Propagation Neural Network (BPNN) classifier using a new feature set from frequency-domain features. The experiments on various categories illustrate that the results of recognition are significant and effective.

General Terms
Feature Extraction, Pattern Classification.

Keywords
Spectral crest, Spectral decrease, Spectral slope, Spectral skewness, Spectral Flatness, Back propagation neural network (BPNN).

1. INTRODUCTION
Sound event recognition is attracting a growing popularity recently in the field of acoustic signal analysis. Not only because it bears great interest for application in multimedia search based on sound, but it is also one of the most important key components to analyze environments, e.g., in surveillance, monitoring of people in need of care, or detecting, localizing, tracking and classifying sources of military interest in real time. Obviously, there is also great benefit for humanoid and general robots, such as the one introduced in for kitchen tasks, to better understand their acoustic environment. Finally, there is hope to better recognize and enhance speech and music, once the sound type of disturbance can be identified [10].

This work consider the task of recognizing environment sounds for the understanding of a scene (or context) surrounding an audio sensor. Here auditory scenes refer to a location with different acoustic characteristics such as a coffee shop, park or quiet hallway. Consider, for example, applications in robotic navigation and obstacle detection, assistive robots, surveillance, and other mobile device based services. Many of these systems are dominantly vision-based. When being employed to understand unstructured environments, their robustness or utility will be lost if visual information is compromised or totally absent. Audio data could be easily acquired, in spite of challenging external conditions such as poor lighting or visual obstruction, and is relatively cheap to store and compute than visual signals. To enhance the system's context awareness, it is needed to incorporate and adequately utilize such audio information [2].

2. RELATED WORK
Research in general audio environment recognition has received some interest in the last few years but the activity is much less as compared to that for speech or music. Most work on audio recognition has focused primarily on speech and music. Less attention has been paid to the challenges and opportunities for using audio to characterize unstructured environments. Unlike speech and music, which have formantic structures and harmonic structures, environmental sounds are considered unstructured since they are variably composed from different sound sources. Applications include those in the domain of wearable and context-aware applications [23], [24]. Unstructured environment characterization is still in its infancy. Most research in environmental sounds has centered mostly on recognition of specific events or sounds [25].

Many previous efforts utilize a high dimension feature vector to represent audio signals [2], [4]. It shows that a high dimension feature set for classification does not always produce good performance. This in turn leads to the issue of selecting an optimal subset of features from a larger set of possible features to yield the most effective subset. It is with these findings that motivated us to look for a more effective approach for representing environmental sounds. Investigations are done in ways of extracting features using matching pursuit and Mel frequency cepstral coefficients (MFCCs) for unstructured sounds in [1]. The features of choice for most audio recognition systems typically rely mostly on the use of MFCC. The filter banks for MFCC are based on the human auditory system and have been shown to work particularly well for structured sounds, like speech and music, but their performance degrades in the presence of noise. MFCC features are modelled based on the shape of the overall spectrum, making it more favourable for modelling single sound sources. Environmental sounds, for example, contain a large variety of sounds, which may include components with strong temporal domain signatures, such as chirpings of insects and sounds of rain. These sounds are in fact noise-like with a broad flat spectrum and are not effectively modeled by MFCCs.

3. OUTLINE OF THE WORK

3.1 Audio event analysis

The first step in building a recognition system for auditory environment was to investigate on techniques for developing an event classification system using audio features. In this work, study was performed by first collecting real world audio from the web which provides a large amount and variety of sound events from real life and then building a classifier to discriminate different environments, which allows us to explore and investigate on suitable features and the feasibility of designing an automatic environment recognition system using audio information. Therefore, this paper proposes a novel feature extraction method that utilizes a small set of frequency-domain features.

3.2 Acoustic features

The purpose of feature extraction is to extract useful discriminative information from the waveform which will result in a compact set of feature vectors [8]. Experimentation are done with many different types of feature for classification of auditory events. Environmental sounds in general are unstructured data comprising of contributions from a variety of sources, and unlike music or speech, no assumptions can be made about predictable repetitions nor harmonic structure in the signal. Due to the inherent diverse nature, there are many features that can be used, or are needed, to describe audio signals. The appropriate choice of these features is crucial in building a robust recognition system. A considerable number of audio features are used in this project from frequency-domain (spectral).

3.2.1 FREQUENCY-DOMAIN FEATURES

3.2.1.1 Spectral Skewness

Spectral skewness is a measure of the asymmetry of the probability distribution of a real-valued random variable that in this context is the spectrum of the signal. For a sample of N values forming a frame, the skewness is:

$$Skewness = \frac{m_3}{m_2^{3/2}} = \frac{\frac{1}{N}\sum_{n=0}^{N-1}(x(n)-\overline{x})^3}{\left(\frac{1}{N}\sum_{n=0}^{N-1}(x(n)-\overline{x})^2\right)^{3/2}} \quad (1)$$

In this equation (1) where $\overline{x}$ represents the mean of the magnitudes, m_3 is the sample third central moment, and m_2 is the sample variance.

3.2.1.2 Spectral Decrease

Spectral decrease also represents the amount of decreasing of the spectral amplitude. This formulation comes from perceptual studies and it is supposed to be more correlated to human perception. The formula is:

$$Decrease = \frac{1}{\sum_{n=1}^{N-1} x(n)} \cdot \sum_{n=1}^{N-1} \frac{x(n)-x(0)}{N-1} \quad (2)$$

In this equation (2) where x(n) represents the weighted frequency value or magnitude of bin number n.

3.2.1.3 Spectral Slope

The spectral slope represents the amount of spectral energy decrease as a function of frequency. It assumes that the amplitude spectrum follows a linear model:

$$A(k) = mk + b \quad (3)$$

The slope m is computed by linear regression.

$$m = \frac{\frac{K}{2}\sum_{k=0}^{\frac{K}{2}-1} kA(k) - \sum_{k=0}^{\frac{K}{2}-1} k \sum_{k=0}^{\frac{K}{2}-1} A(k)}{\frac{K}{2}\sum_{k=0}^{\frac{K}{2}-1} k^2 - \left(\sum_{k=0}^{\frac{K}{2}-1} k\right)} \quad (4)$$

In this equation (3), (4) where K is the total number of frequency values, A(k) is the spectral magnitude with frequency index k.

3.2.1.4 Spectral Crest

Spectral crest factor indicate how flat or "peaky" the power spectral density is in a given subband. The Spectral Crest Factor or peak-to-average ratio is a measurement of a waveform, calculated from the peak amplitude of the waveform divided by the mean value of the waveform.

$$Crest\ Factor = \frac{|x_n(i)|_{peak}}{\sqrt{\frac{\sum_{i=1}^{N} x_n{}^2(i)}{N}}} \quad (5)$$

In this equation (5) where N is the frame length, and $x_n[i]$ represents spectral amplitude of the i th sample in the nth frame

3.2.1.5 Spectral Flatness (SF)

Spectral Flatness is a measure of distribution of spectral power in an audio spectrum. The spectral flatness is calculated by dividing the geometric mean of the power spectrum by the arithmetic mean of the power spectrum. The spectral flatness used is measured across the whole band. The formula is:

$$\mathrm{SF} = \frac{\sqrt[N]{\prod_{n=0}^{N-1} x(n)}}{\frac{\sum_{n=0}^{N-1} x(n)}{N}} \quad (6)$$

In this equation (6) where $x(n)$ represents the magnitude of bin number n of the power spectrum with a frame length of N.

3.3 Neural network

Neural networks which have been widely used in image and signal processing are very effective for solving multiple class classification problems. Neural networks are composed of simple elements operating in parallel. These elements are inspired by biological nervous systems. As in nature, the connections between elements largely determine the network function. Neural networks have been trained to perform complex functions in various fields, including pattern recognition, identification, classification, speech, vision, and control systems. Neural networks can also be trained to solve problems that are difficult for conventional computers or human beings [6].

The recognition performance of the neural network will highly depend on the structure of the network and training algorithm. It is consists of three layers forward structure that has hidden layer between input layer and output layer interconnected by links that contains weights.

3.4 Back propagation neural network (BPNN) for audio classification

In this system, feed-forward neural network with back propagation learning algorithm is used. Two different components make up the construct of phases: one is the feed forward phase in which the external input information at the input nodes is propagated forward to compute the output information signal at the output unit, and a backward phase in which modification to the connection strengths is made based on the differences between the computed and observed information signals at the outputs units [16].Figure 1, shows the Back propagation neural network frame [17].

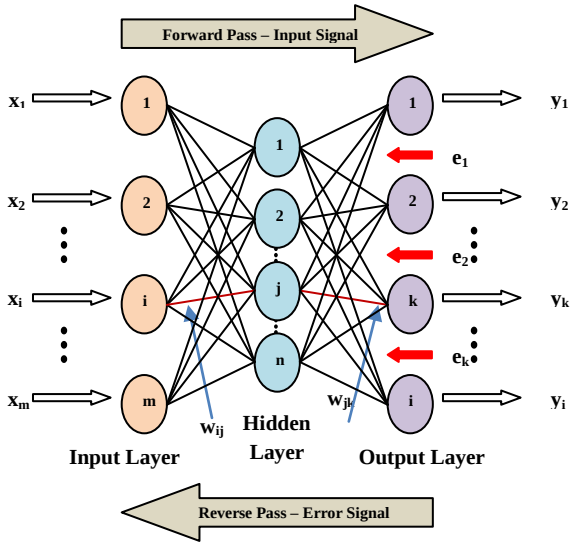

Fig 1: Back propagation neural network

The input for a certain neuron might either be very large or negative. It is to be noted that this is generally not desirable. In order to avoid large or negative values and to introduce nonlinearity in the model, the neuron's input undergo an nonlinear transformation to produce an output:

$$(out)_m = f\,(input)_m$$

Where f is a transfer function and $(input)_m$ is the value calculated in previous neuron. The integrated signal is transformed to activation via a transfer function such as the sigmoidal function. Sigmoidal function is a continuous activation function, designed to respond relative to the amount of excitation received [11]. It is the most widely used function in various BPNN applications.

4. EXPERIMENTAL ANALYSIS AND RESULTS

This section deals with experimental implementations and evaluations of results of the proposed system, both on audio event detection and on semantic inference of the auditory context.

4.1 Database

The database for the experiments contains 1000 samples which are taken from AURORA database. The recordings are categorized into general classes according to common characteristics of the scenes (220 kitchen noises, 180 living room noises, 210 laundry sounds, 230 meeting sounds, 160 office sounds) and events (Pan boiling, steel plate ,music player, paper scrap ,washing machine, flush,overlapped speech, footsteps, typewriter, dust bin, etc.). The categorization of the scenes was somewhat ambiguous; some of the recordings are associated with more than one higher-level class. The recordings are manually labelled and are separated into 2-second, 3-second and 5-second fragments. Every sound signal was stored with some properties that are also the initial conditions and criteria for the well-functioning of the algorithm. The sample database is split into training sets and test sets. The dataset randomly contains 80% sounds of each class for the training set. The remaining 20% sounds form the test set. Thus it has taken different proportion of samples based on class dependency in each category as shown in table 1.

Table 1: ACOUSTIC DATABASE DESCRIPTOR

Context	Total amount of database
Kitchen	22 %
Living Room	18 %
Laundry	21 %
Meeting	23 %
Office	16 %

4.2 Experimental evaluation

This subsection performs an analysis on the performance of the proposed method. Figure 2, shows the block diagram of this sound recognition system. The Environmental scene classification system comprises of three main modules: 1) Pre-processing, 2) Feature extraction and 3) Final classifier.

4.2.1 Pre-processing

To extract the features from the acoustic signal, the signal must be pre-processed and divided into successive windows or analysis frames. Throughout this work, a sampling rate of 16 kHz, 16 bit monophonic, pulse code modulation (PCM) format in wave audio is adopted. Environmental audio signal which is recorded using multi-microphone setting is pre-processed before extracting features. This involves normalization of audio waveform, pre-emphasis and windowing of the frame. The process of pre-emphasis provides high frequency emphasis and windowing which reduces the effect of discontinuity at the ends of each frame in the audio. The training data is segmented into fixed-length and overlapping frames (in this experiments, 20 ms frames with 10 ms overlapping is utilized). When neighboring frames are overlapped, the spectral characteristics of audio content can be taken into consideration in the training process.

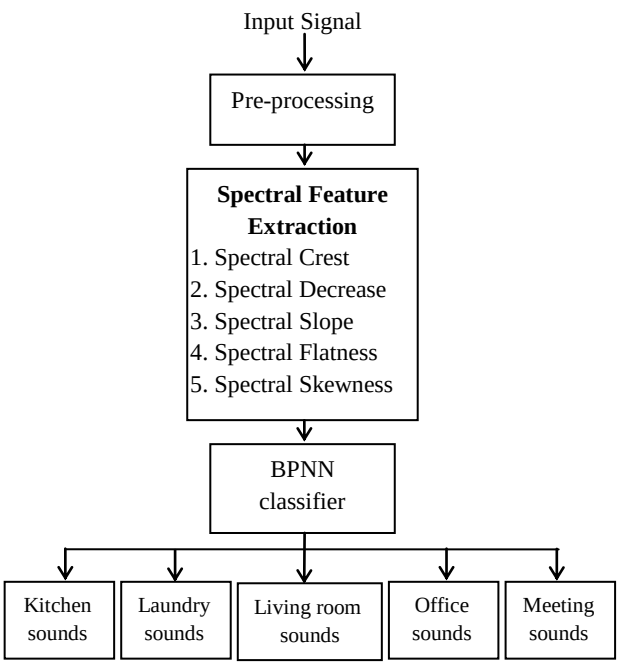

Fig 2: Block diagram for sound recognition system

4.2.2 Feature extraction

Extracting the right feature set for environmental sounds is the key to effective performance. A variety of features in different domain have been proposed for audio recognition, but the vast majority of the past work utilizes features that are well-known for structured data, such as speech and music, and assumes this association will transfer naturally well to unstructured sounds. Since environmental sounds may differ significantly from speech, additionally features that address the possibility of high non-stationary of sounds are considered.

In this paper, proposed a novel method based on spectral domain to analyse environmental sounds for their feature extraction. Thus a 5 dimensional feature matrix is obtained for each frame. Features such as crest factor, flatness, skewness, decrease and slope are calculated which forms the feature matrix. The proposed method shows that these features are capable of effectively representing sounds that originate from different sources and different environments. Thus resulting in an representation that is flexible, yet intuitive and physically interpretable.

4.2.3 Modeling using BPNN

For modeling back propagation neural network classifier is used here to discriminate various events. Classification parameters are calculated using BPNN learning. The training process analyzes audio training data to find an optimal way to classify audio frames into their respective classes. The training data should be sufficient to be statistically significant. The BPNN learning algorithm is applied to produce the classification parameters according to calculated features. The derived classification parameters are used to classify the context of the audio data. The classification results for the proposed feature set are shown in Figure 3 for various sample durations. From the results, it is observed that the overall classification accuracy is high for 3-second samples when compared to other duration.

Table 2: Recognition matrix for 15 hidden neurons

	2 second	3 second	5 second
Recognition Accuracy	79%	91.7%	86%

To determine the performance of BPNN, results for various numbers of neurons in the hidden layer is examined. Using the same settings as the rest of the experiments, performance are examined for hidden neurons of 5, 10, 15 and 20 and used the same number of neurons for each environment type. The overall recognition rates are plotted in figure 4. It is shown that the classification performance peaks around fifteen and the performance slowly degrades as the number of neurons varies. The highest recognition rate for each class across the number of mixtures was obtained with 15 neurons.

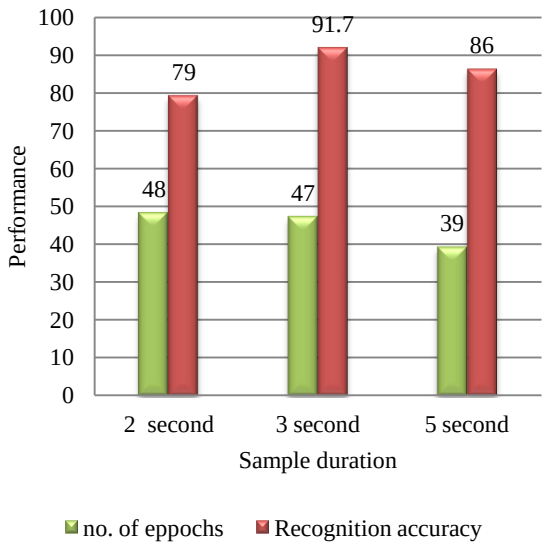

Fig 3: Recognition chart for different sound samples for 15 hidden neurons

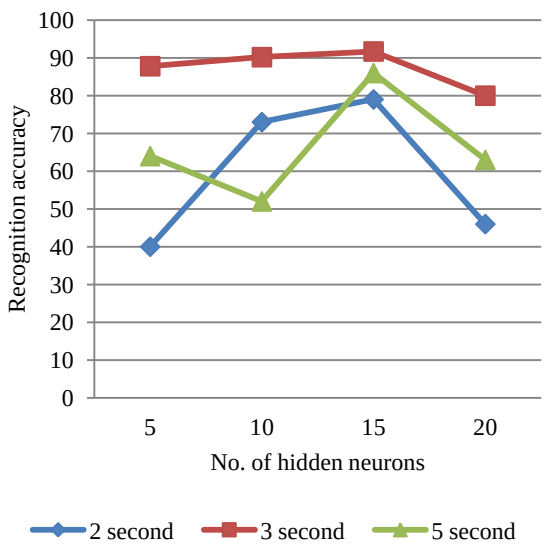

Fig 4: Recognition chart for different hidden neurons

4.2.4 Confusion matrix

Results presented in charts are averaged values from all trials together. To further understand the classification performance, results are shown in the form of a confusion matrix, which allows us to observe the degree of confusion among different classes. The confusion matrix given in Table 3 is built from a single arbitrary trial, constructed by applying the classifier to the test set and displaying the number of correctly/incorrectly classified items. The rows of the matrix denote the environment classes that are attempted to classify, and the columns depict classified results. It is seen from Table 3 that Laundry, and Meeting was more often misclassified than the rest.

Table 3: Confusion Matrix For 5-Class Classification Using Spectral Features With BPNN

Accuracy %	Kitchen	Living Room	Laundry	Meeting	Office
Kitchen	206	3	4	2	5
Living Room	3	164	6	7	0
Laundry	4	6	199	0	1
Meeting	2	7	0	218	3
Office	5	0	1	3	151
Overall Accuracy: 91.7%					

5. CONCLUSION AND FUTURE WORK

For recognition of all the 5-classes, it is shown that the BPNN did achieve significant overall recognition for every independent event. By varying the parameters of the learning rate and hidden neurons (Table 2) that managed to increase the average recognition rate to 91.7%. The experimental results show that the proposed method with BPNN is very effective recognition method and can accomplish event recognition in a short time and achieve a satisfactory recognition rate of 91.7%.

The need for creation of quality Environment sound recognition is also eminent in nowadays. Despite some small setbacks, environmental acoustic assessment has become an integral part of pattern classification, which is continually being improved for posterity.

6. REFERENCES

[1] Selina Chu, Shrikanth Narayanan and C.-C. Jay Kuo, “Environmental Sound Recognition With Time-Frequency Audio Features”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 17, No. 17, NO. 6, AUGUST 2009.

[2] Behnaz Ghoraani, and Sridhar Krishnan, “Time–Frequency Matrix Feature Extraction and Classification of Environmental Audio Signals”, IEEE Transactions on Audio, Speech, and Language Processing, Vol. 19, No. 7, September 2011.

[3] Jia-Ching Wang, Hsiao-Ping Lee, Jhing-Fa Wang, and Cai-Bei Lin, “Robust Environmental Sound Recognition for Home Automation”, IEEE Transactions on Automation Science and Engineering, Vol. 5, No. 1, January 2008.

[4] Asma Rabaoui, Manuel Davy, Stéphane Rossignol, and Noureddine Ellouze, “Using One-Class SVMs and Wavelets for Audio Surveillance”, IEEE Transactions on Information Forensics and Security, Vol. 3, No. 4, December 2008.

[5] Regunathan Radhakrishnan, Ajay Divakaran and Paris Smaragdis, “Audio Analysis for Surveillance Applications”, 2005 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics.

[6] Supreet Kaur, and Er. Ranjeet Kaur Sandhu, “A Survey on Enhanced Human Identification Using Gait Recognition based on Neural Network And Support Vector Machine”, International Journal of Application or Innovation in Engineering & Management (IJAIEM), Web Site: www.ijaiem.org Email: editor@ijaiem.org, editorijaiem@gmail.com, Vol. 2, Issue 6, June 2013 ISSN 2319 – 4847 .

[7] Saurabh Karsoliya , “Approximating Number of Hidden layer neurons in Multiple Hidden Layer BPNN Architecture”, International Journal of Engineering Trends and Technology Volume3, Issue 6- 2012, ISSN: 2231-5381.

[8] Ling Ma, Dan Smith and Ben Milner , “Environmental Noise Classification for Context-Aware Applications”, School of Computing Sciences, University of East Anglia Norwich, NR4 7TJ, UK, {ling.ma, dan.smith, b.milner}@uea.ac.uk .

[9] Michael Cowling and Renate Sitte, “Analysis of Speech Recognition Techniques for use in a Non-Speech Sound Recognition System”, Griffith University Faculty of Engineering & Information Technology, Gold Coast, Qld, Australia 9726.

[10] Zixing Zhang and Bj¨orn Schuller, “Semi-Supervised Learning Helps in Sound Event Classification”, Institute for Human-Machine Communication, Technische Universit¨at M¨unchen, Germany, zixing.zhang|schuller)@tum.de .

[11] Victoria Felker, Mustaque Hossain, Yacoub Najjar , Richard Barezinsky , “Modeling the Roughness of Kansas PCC Pavements: A Dynamic ANN Approach”, For Presentation at the 82nd Annual Meeting of the Transportation Research Board January 12-16, 2003 and Publication in the Transportation Research Record.

[12] Selina Chu, “Unstructured Audio Classification for Environment Recognition”, Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence (2008).

[13] Dr. Yadana Thein , San Su Su Yee, “High Accuracy Myanmar Handwritten Character Recognition using Hybrid approach through MICR and Neural Network” , IJCSI International Journal of Computer Science Issues, Vol. 7, Issue 6, November 2010 ISSN (Online): 1694-0814.

[14] David Gerhard , “Audio Signal Classification: History and Current Techniques”, Technical Report TR-CS 2003-07 November, 2003, ISSN 0828-3494, ISBN 0 7731 0456 9.

[15] Dalibor Mitrovi_, Matthias Zeppelzauer, Horst Eidenberger, ”On Feature Selection in Environmental Sound Recognition”, 51st International Symposium ELMAR-2009, 28-30 September 2009, Zadar, Croatia.

[16] Jian-Da Wu, Yi-Jang Tsai, “Speaker identification system using empirical mode decomposition and an artificial neural network”, Expert Systems with Applications 38 (2011) 6112–6117.

[17] T.B.Adam, Md Salam, “Spoken English Alphabet Recognition with Mel Frequency Cepstral Coefficients and Back Propagation Neural Networks”, International Journal of Computer Applications (0975 – 8887) Vol. 42– No.12, March 2012.

[18] Giovanni De Poli, "From audio to content", Chapter 4, Copyright 2006.

[19] Wei Chu, "Auditory-Based Noise-Robust Audio Classification Algorithms", Dept. of Electrical & Computer Engineering, McGill University Montreal, Canada, September 2008, thesis submitted to McGill University for fulfillment for the degree of Doctor of Philosophy.

[20] Eric Allamanche, J¨urgen, Oliver Hellmuth, Thorsten Kastner, Christian Ertel, "A Multiple Feature Model for Musical Similarity Retrieval" , Fraunhofer Institute Integrierte Schaltungen, IIS Am Wolfsmantel 33 D-91058 Erlangen, Germany.

[21] Renato Eduardo Silva Panda, "Automatic Mood Tracking in Audio Music", Master in Informatics Engineering, M.Sc. Thesis, panda@student.dei.uc.pt Thesis Supervisor: Professor Rui Pedro Paiva July, 2010.

[22] Federico Avanzini Riccardo Levorato Emanuele Menegatti, "GMM Classification of Environmental Sounds for Surveillance Applications", University of Padova Department of Information Engineering, Master's Degree in Computer Engineering (Laurea Magistrale in Ingegneria Informatica) Graduation Date: Padova, October 26th, 2010.

[23] Kyuwoong Hwang and Soo-Young Lee, "Environmental Audio Scene and Activity Recognition through Mobile-Based Crowdsourcing", IEEE Transactions on Consumer Electronics, Vol. 58, No. 2, May 2012.

[24] Woo-Hyun Choi, Seung-Il Kim, Min-Seok Keum and David K. Han, Hanseok Ko, "Acoustic and Visual Signal Based Context Awareness System for Mobile Application", IEEE Transactions on Consumer Electronics, Vol. 57, No. 2, May 2011.

[25] J.D. Krijnders, M.E. Niessen, T.C. Andringa, "Sound event recognition through expectancy-based evaluation of signal-driven hypotheses", ELSEVIER, Pattern Recognition Letters 31 (2010) 1552–1559.

[26] Laura Enflo, "Spectral Tilt Used for Automatic Detection of Prominence: A Comparison between Electroglottography and Audio", (Term Paper for Machine Learning, a GSLT course 2009).