Proceedings of the 14th IFAC Symposium on
Information Control Problems in Manufacturing
Bucharest, Romania, May 23-25, 2012

An Environmental Sound Source Classification System Based on Mel-Frequency Cepstral Coefficients and Gaussian Mixture Models

Guanghu Shen*, Quang Nguyen* and JongSuk Choi*

*Center for Bionics, Korea Institute of Science and Technology, Seoul, Korea
(Tel: +82-2-958-5618; e-mail:{khshin, qvnguyen, cjs}@kist.re.kr).

Abstract: This paper proposed a study of a sound source classification system that has been developed for detecting and identifying the detected sound events in real environments. The proposed system was based on a pattern recognition approach using Gaussian mixture models and Mel-Frequency Cepstral Coefficients (MFCCs) features. We considered eight types of basic sound sources and an external sound. To make the system robust to various types of sound sources, we designed a tree of reference sound models for classification, in which especially generated total three of GMMs for external sounds according to different characteristics of frequency distributions. The performance of the proposed system, evaluated in terms of percent classification, indicated an averaged accuracy of 91.36% for off-line test. Finally, in on-line test our proposed system also showed a good and stable performance in real environments.

Keywords: Pattern recognition, Impulsive sound detection, Sound classification, MFCCs, Gaussian Mixture models

1. INTRODUCTION

Automatic environmental sound source classification has become a very active subject of research during last decades, since it can be directly or indirectly implemented into a very wide area of topics, including speech recognition, pattern identification, and context-aware applications. Especially, a few context-aware applications have attempted to use environmental sound sources. For example, in noise monitoring systems, classification of environmental noises has been provided to help in controlling noise pollutions. And the use of a sound event detection and classification system can offer concrete potentialities for surveillance and security applications, by contributing to alarm triggering or validation. Furthermore, these functionalities can also be used in portable tele-assistive devices, to inform disabled and elderly persons affected in their hearing capabilities about relevant environmental sounds (warning signals, etc.) [1].

Many different methods and algorithms are developed for identifying noise sources. While one group of studies focuses on the extraction of the feature parameters such as Linear Prediction Coding (LPC), Perceptual Linear Predictive (PLP) and Mel-Frequency Cepstral Coefficients (MFCCs), etc; the other group of studies focus on the classification techniques such as Hidden Markov Models (HMMs), statistical pattern recognition systems, Artificial Neural Networks (ANNs), fuzzy logic systems, etc. [2]

Typically, F. Beritelli et al. [3] proposed a sound classification system based on MFCCs feature and ANNs classifier, and it showed an accuracy of 85-92% for 10 kinds of environmental noise sources. L. Ma et al. [4] proposed a system based on MFCCs feature and HMM classifier, and it showed a good classification performance for 11 kinds of environmental noise sources, especially the system showed the highest accuracy of 92.27% when the number of states in HMM was set between 11 to 15.

A. Dufaux et al. [1] compared the performance of GMMs and HMMs classifiers which use the energy of uniform-spectral bands as feature of 6 types of impact sound sources, as a result HMMs showed a better performance than GMMs by achieving classification accuracies of 91% and 86%, respectively.

L. Couvreur et al. [5] presented a classification system based on ANNs coupled with HMMs and PLP feature, and it showed a classification accuracy of 85% for urban environmental noise sources such as the scooter engine and horn signals.

S. Ntalampiras et al. [6] presented a two-stage GMM-HMM classification system based on MFCCs or MPEG-7 feature set to classify four types of mechanic sound sources and four kinds of non-mechanic sound sources. For the classification process, the first stage classifies sound into two categories (mechanic and non-mechanic) using GMMs classifier while the second stage completes the rest of the classification process producing the leaf-class using HMMs classifier. As a result, MPEG-7 and MFCCs showed the highest accuracy of 81% and 78% in each experiment, respectively. However, the problem of the two-stage classification approach is that it is impossible to handle when a classification error was occurred in the first-stage of classification.

R. Annies et al. [7] focused on classification footstep sounds including five types of shoes, and six types floors,

978-3-902661-98-2/12/$20.00 © 2012 IFAC

10.3182/20120523-3-RO-2023.00251

which used the feature based on Gamma-tone Auditory Filter-bank. As a result, SVM (Support Vector Machines) classifier showed an accuracy of 95%, while HMMs classifier showed an accuracy of 87%.

Finally, A. Rabaoui [8] presented a HMM-based classification system for sound sources in real traffic environments such as car, truck, train etc., and compared the classification performances of different features. As a result, MFCCs, LPC and PLP showed the best performance of 93%, 88% and 93%, respectively.

Based on the analysis of the existing sound classification systems above, we found that most studies showed a good performance in experiments conducted in lab environments. However, we consider that those systems are difficult to implement in real environments, since there exist various types of sound sources. In this study, we aimed to real-time application of sound classification, especially robust in real environments. We proposed a classification system based on MFCCs and GMMs, which is not only considering the computational cost of algorithms but also the classification performance in real environments.

This paper is organized as follows. An introduction to our proposed sound source classification system is described in section 2. In section 3, we evaluate the performance of our system by conducting both off-line and on-line experiments. Finally, we conclude this work in section 4.

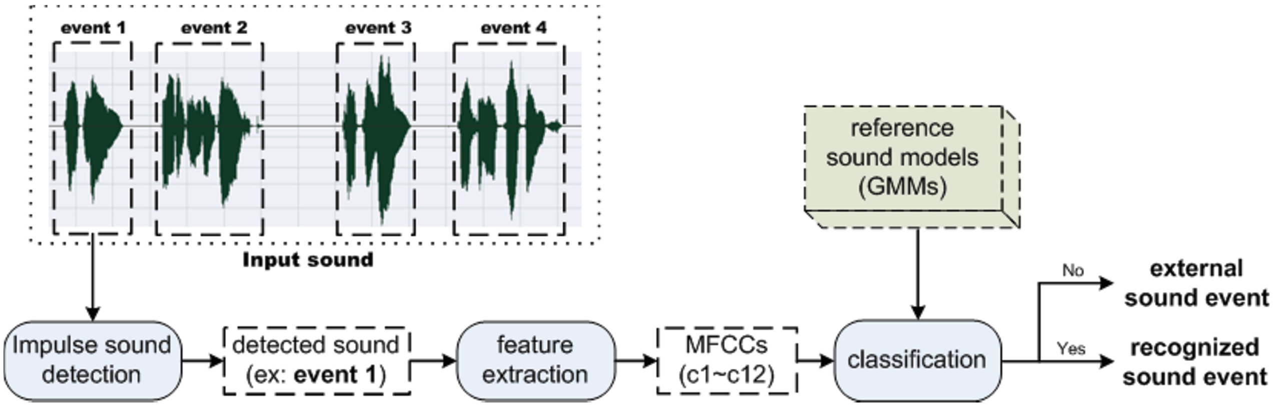


Fig. 1. The block diagram of the proposed sound source classification system.

2. PROPOSED SOUND SOURCE CLASSIFICATION SYSTEM

The block diagram of the proposed sound source classification system is shown in Fig. 1. It starts with detection of sound event by the impulse sound detection module, extraction of feature parameters for each sound event pause is then activated. Finally, the detected sound is recognized after comparison with reference sound models. Each block will be described in detail in following sections.

2.1 Impulse Sound Detection

In real-life environments, there exist various types of sound sources such as sound with low frequencies, sound with middle frequencies, or sound with high frequencies. In this paper, we calculate short-time energy (STE) in frequency domain as (1),

$$P(n) = 10\log 10(\sum X^2(n,f)) \tag{1}$$

where $X(n,f)$ is the short-time Fourier transform of $x(n)$, n is the frame index, f is the frequency bin.

Decision is made by comparison of the value of STE with a threshold value:

$$VAD(n) = \begin{cases} 1 & if\ P(n) - P_{noise}(n) > TH \\ 0 & otherwise \end{cases} \tag{2}$$

where $P_{noise}(n), TH$ is the power of background noise at the n-th frame, the threshold value (i.e., 3dB), respectively. In other words, the current frame is decided as sound event pause if the difference of values of STE between the input frame and the background noise is greater than the threshold (*TH*), otherwise the current input frame is decided as no event.

When the background noise changes, the threshold should be adaptively change. Here we can estimate the background noise level by smoothing it during noisy frame frames

$$P_{noise}(n+1) = \begin{cases} \alpha_1 P_{noise}(n) + (1-\alpha_1)P(n) & if\ VAD(n) = 0 \\ \alpha_2 P_{noise}(n) + (1-\alpha_2)P(n) & otherwise \end{cases} \tag{3}$$

where α_1 , α_2 is the estimation factor 0.95, 0.99, respectively.

2.2 Feature Extraction

The process of extracting MFCCs feature parameters is depicted in Fig. 2 [4]. Feature extraction begins by estimating the magnitude spectrum of each frame from the input sound data (Discrete Fourier Transform, DFT). The magnitude spectrum is then non-linearly quantised using a mel-scale filter bank (Mel Filter Bank, MFB) which models the psychoacoustic properties of the human ear. Finally, the logarithm (Log) is taken to outputs of filter-bank bands and then a discrete cosine transform (DCT) is applied as follows,

$$c_n = \text{[redacted]} \sum_{j=1}^{N} m_j \cos(\frac{\pi n}{N}(j-0.5)) \tag{4}$$

where c_n is MFCCs feature vectors of the *n*-th frame, N is the number of filter bank channels and m_j the output of the *j*-th channel of mel-scale filter banks.

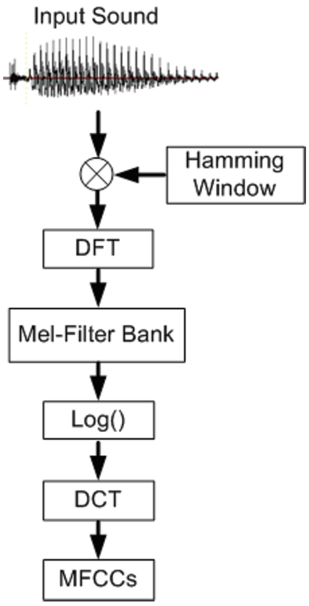


Fig. 2. Flow chart of feature extraction

2.3 Gaussian Mixture Models Description

A Gaussian mixture density is a weighted sum of *M* component densities [9], as depicted in Fig. 3 and is given by the equation

$$p(\vec{x} \mid \lambda) = \sum_{i=1}^{M} p_i b_i(\vec{x}) \tag{5}$$

where $\vec{x}$ is a *D*-dimensional random vector, $b_i(\vec{x})$, $i = 1,...,M$ are the component densities and $p_i, i = 1,...,M$ are the mixture weights. Each component density is a *D*-variate Gaussian function of the form

$$b_i(\vec{x}) = \frac{1}{(2\pi)^{D/2} |\Sigma_i|^{1/2}} \exp\left\{ -\frac{1}{2}(\vec{x} - \vec{\mu}_i)'\Sigma_i^{-1}(\vec{x} - \vec{\mu}_i) \right\} \tag{6}$$

With mean vector $\vec{\mu}_i$ and covariance matrix Σ_i. The mixture weights satisfy the constraint that $\sum_{i=1}^{M} p_i = 1$.

The complete Gaussian mixture density is parameterized by the mean vectors, covariance matrices and mixture weights from all component densities. These parameters are collectively represented by the notation

$$\lambda = \{p_i, \vec{\mu}_i, \Sigma_i\} \quad i = 1,...,M \tag{7}$$

For sound source classification, each sound source is represented by a GMM and is referred to by its model λ.

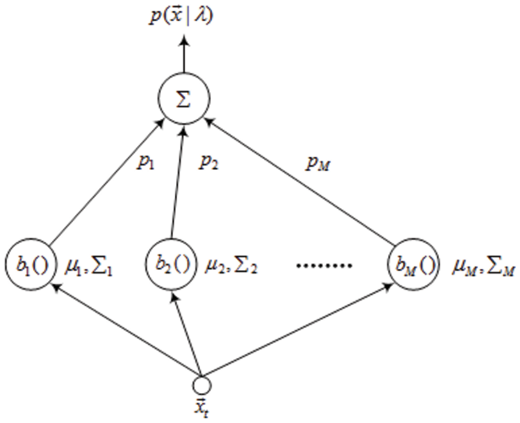


Fig. 3. Depiction of an *M* component Gaussian mixture density

2.4 Maximum Likelihood Parameter Estimation

For sound source classification, a group of sound sources $S = \{1, 2, ..., S\}$ is represented by GMMs ($\lambda_1, \lambda_2, ..., \lambda_S$). The objective is to find the sound model which has the maximum *a posteriori* probability for a given observation sequence. Formally,

$$\hat{S} = \arg\max_{1 \le k \le S} p(\lambda_k \mid X) = \arg\max_{1 \le k \le S} \frac{p(X \mid \lambda_k) p(\lambda_k)}{p(X)} \tag{8}$$

where the second equation is due to *Bayes'* rule. Assuming equally likely sound sources (i.e., $p(\lambda_k) = 1/S$) and nothing that $p(X)$ is the same for all sound models, the classification rule simplifiers to

$$\hat{S} = \arg\max_{1 \le k \le S} p(X \mid \lambda_k) \tag{9}$$

Using logarithms and the independence between observations, the sound classification system computes

$$\hat{S} = \arg\max_{1 \le k \le S} \sum_{t=1}^{T} \log p(\vec{x}_t \mid \lambda_k) \tag{10}$$

in which $p(\vec{x}_t \mid \lambda_k)$ is given in (5), and t is the time index.

3. EXPERIMENTAL EVALUATION

3.1 Database Description

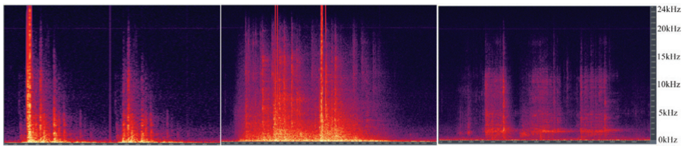
To test the performance of the proposed system, we collected data from various resources including the BBC Sound Effects

Library [10] and some recordings found on the internet. As shown in Table 1, nine categories were organized containing different numbers of sound data. To conduct task-independent classification experiments, we divided the number of sound data into two groups (i.e., training set and test set) by the ratio of 2:1. In detail, the classes are alarm, door bell, door knock, explosion, footstep, glass smash, door open & close (general, metal, sliding), scream (man, woman & child), and an external sound.

Table 1. Sound database used in experiments

sound types		train	test
alarm		19	9
door bell		11	6
door knock		27	14
door open&close	general	17	8
	metal	78	40
	sliding	59	29
explosion		15	7
footstep		14	5
glass smash		48	24
scream	man	12	6
	woman&child	14	7
external sound		100	16

In addition, we divided door open&close sound into three subgroups (i.e., general, metal and sliding). Since the frequency characteristic of the sound of door open&close is very different depending on the material of the door or the way of opening (such as sliding). Fig. 4 shows the spectrogram of three samples of door open&close sound source, Fig. 4 (1) is for the sound of general type of door such as plastic door, wood door, etc. In this case, we found that most energies are distributed in low frequency bands (ex., under 6 kHz). On the other hand, Fig. 4 (2) is for the sound of metal type of door. In this case, we found that most energies are distributed in higher frequency bands (ex., under 17 kHz) comparing with the case of general type. Finally, Fig. 4 (3) is for the sound of sliding door. This sound is more like white noise, its energies are approximately equally spreaded in whole frequency bands.



(1) general type (2) metal type (3) sliding type

Fig. 4. Spectrogram of three samples of door open&close sound source

In a similar way, the scream sound is divided into two subgroups. One subgroup is man's scream voice, the other is woman's & child's scream voice. Since those two subgroups have large differences in fundamental frequency.

All sound data were 48 kHz sampling frequency, 16 bits quantization. Each data was put into a preprocessing based on a process of pre-emphasis filtering with a coefficient of $\alpha = 0.95$; fragmentation, in which the data was subdivided into 20 ms frames with a 15 ms overlap between consecutive frames; and then the application of a Hamming window to ensure smoothing. Finally, a 12-dimensional MFCCs was extracted from each frame.

3.2 Classification Schema for External Sound

In particular, since there are a various types of sound sources in real environments, in this work we took into account a total of nine types of sound sources, i.e., eight types of basic sound sources (alarm, door bell, door knock, explosion, footstep, glass smash, scream, door open & close) and an external sound. Here, the external sound means the other sound sources generated popularly in real-life environments but excepts basic types of sound sources, such as animal sounds (dogs, cats, etc.), nature sounds (rain, wind, etc.) and human sounds (laugh, applause, etc.).

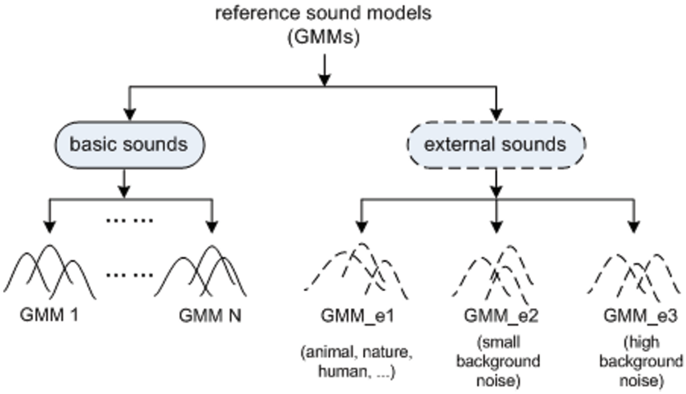


Fig. 5. Tree of reference sound models (GMMs)

In order to make a robust sound classification for various types of sound sources in real environments, we designed a tree of reference sound models (i.e., GMMs) as shown in Fig. 5. In detail, we generated a total of 11 GMMs for basic types of sound sources by using training data of each type of sound source (left part of Fig. 5). Here, according to the analysis in section 3.1, we generated a GMM for each subgroup in both cases of door open&close sound and scream sound. By this way we could expect to increase the accuracy of classification for basic types of sound sources because of its improved discriminant of acoustic models by adding exact GMMs for each subgroup.

On the other hand, we generated a total of three GMMs for external sound by using three groups of dataset, i.e., other sounds (such as animal, natural and human sounds, etc., but except basic sound types), a set of background sounds with small levels, and a set of background sounds with high levels (right part of Fig. 5). If one external sound event happened, our proposed system is more easily classified as the type of external sound comparing with the case of only one (or without) GMMs for external sound. As a result, by using our proposed classification schema for external sound, we can not only keep high classification accuracies for basic types of sound sources but also improve the stability of classification for external sounds. We consider that it might be an

Table 2. Comparison of classification accuracies of different sound sources

Reference Test dat	alarm	door _bell	door _knock	door _open&close	explosion	footstep	glass _smash	scream	**external _sound**
alarm	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
door_bell	0.0%	83.3%	0.0%	16.7%	0.0%	0.0%	0.0%	0.0%	0.0%
door_knock	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
door_open&close	0.0%	0.0%	0.0%	92.2%	2.6%	0.0%	2.6%	0.0%	2.6%
explosion	0.0%	0.0%	0.0%	42.9%	57.1%	0.0%	0.0%	0.0%	0.0%
footstep	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%	0.0%	0.0%
glass_smash	0.0%	0.0%	0.0%	0.0%	0.0%	4.2%	95.8%	0.0%	0.0%
scream	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	100.0%	0.0%
external_sound	0.0%	0.0%	0.0%	6.3%	0.0%	0.0%	0.0%	0.0%	**93.8%**

Important factor for on-line sound source classification system, since there are various types of sound sources in real-life environments.

3.3 Experimental Results

Table 2 shows the experimental results of so und classification for 9 types of sound sources. The first column in the table is the list of test sound sources, and the first raw in the table is the list of reference GMMs. Each number in diagonal position represents the classification accuracy of each corresponding type of sound source. As a result, alarm, door bell, door knock, door open&close, explosion, footstep, glass smash, scream, and external sound show the accuracy of 100%, 83.3%, 100%, 92.2%, 57.1%, 100%, 95.8%, 100%, and 93.8%, respectively. Finally, we can get the averaged accuracy of all types of test sound data is 91.36%. Thus we can confirm that our proposed sound classification system is efficient in various types of sound sources.

However, we could also find several problems in our proposed system from Table 2. First, door bell sound shows the ratio of misclassification as door open&close is 16.7%. We considered that the possible reason might be that some types of door bell are very similar with door open&close (metal type), since both two types of sound have similar characteristic that most energies are distributed in high frequency bands. In a similar reason, many explosion data were misclassified as door open&close (sliding), since both two types of sound have similar characteristic that most of energies are approximately equally distributed in whole frequency bands.

Another problem of these two cases of misclassifications is that the number of training data of each type is a bit small. Therefore, those problems might be solved if we collect more data for each type of sound source for training the reference GMMs models.

3.4 On-line Sound Classification System

In this work, we developed an on-line sound source classification system with Visual C++ language. Fig. 6 shows the Graphical User Interface (GUI) window when the system runs in real-time. The left part in Fig. 6 shows the type of sound sources by its image, which is registered in advance to the list of reference sound sources. The upper right window is to display the waveform of input sound, and the bottom of the right window is to display the output result of sound classification. When the "Run" button is clicked, the sound classification system will run on real-time. In detail, whenever the impulse sound detection module detects a sound event, the system will extract features of the detected sound event for successive 50~100 frames (Here, we set as 80 frames), and then deliver those features to the classification stage. Finally, the corresponding image of classification result will show in the window. From the on-line sound classification test conducted in various real-life environments, we can confirm that our proposed system runs very well and shows a good and stable classification performance for various types of sound sources.

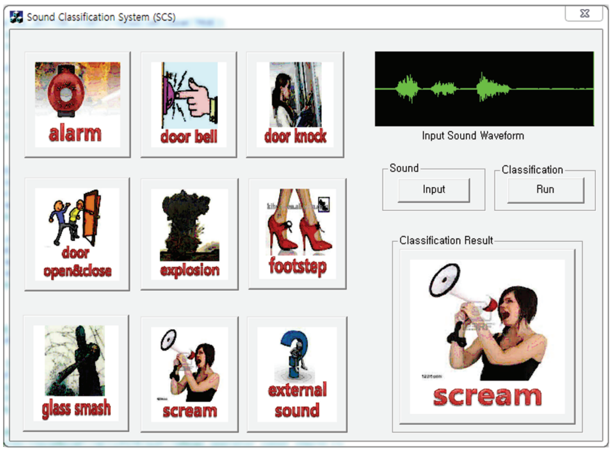


Fig. 6. A GUI window of the proposed on-line sound source classification system

4. CONCLUSIONS

In this paper, we proposed an on-line sound source classification system based on MFCCs feature parameters and GMMs models. To make it robust in various types of sound sources, we designed a tree of reference sound models for classification, in which especially total three of GMMs for external sounds were generated according to different characteristics of frequency distributions. From experimental results, the proposed sound classification system showed an averaged accuracy of 91.36% for off-line test. Finally, to confirm the performance of our system in real applications, we performed various on-line tests in many real-life environments, as a result the system showed a good and stable performance for various types of real sound events.

Research directions for future work include the test of the proposed system to more difficult task (more sound types) and more robust for rejecting misclassified sound events. Besides, the algorithm should be optimized and integrated on a DSP so as to incorporate it in real-life monitoring systems.

REFERENCES

[1] A. Dufaux, L. Besacier, M. Ansorge, and F. Pellandini (2000), "Automatic sound detection and recognition for noisy environment," *Proc. EUSIPC*, pp. 1033-1036, Tampere, FI, September 5-8, 2000.

[2] B. D. Barkana, and I. Saricicek (2010), “Environmental noise source classification using neural networks,” *Proc. ICIT*, pp. 259-263, April 2010.

[3] F. Beritelli, and R. Grasso (2008), “A pattern recognition system for environmental sound classification based on MFCCs and neural networks," *Proc. IEEE ICSPCS*, pp. 1-4, Dec. 2008.

[4] L. Ma, D. Smith, and B. Milner (2003), "Enviromental noise classification for context-aware applications," *DEXA 2003, LNCS 2736*, pp. 360-370.

[5] L. Couvreur, and M. Laniray (2004), "Automatic noise recognition in urban environments based on artificial neural networks and hidden markov models," *Proc. 33rd Inter-noise*, Prague, Czech Republic.

[6] S. Ntalampiras, I. Potamitis, and N. Fakotakis (2008), “Automatic recognition of urban environmental sound events,” Proc. International Association for Pattern Recognition Workshop on Cognitive Information Processing, Santorini, Greece.

[7] R. Annies, E. Martinez, K. Adiloglu, H. Purwins, and K. Obermayer (2007), "Classification schemes for step sounds based on gammatone-filters," Proc. workshop at NIPS.

[8] A. Rabaoui, Z. Lachiri, and N. Ellouze (2004), "Automatic environmental noise recognition," Proc. IEEE ICIT, pp. 1670-1675.

[9] D. A. Reynolds, and R. C. Rose (1995), “Robust text-independent speaker identification using Gaussian mixture speaker models,” IEEE Transactions on Speech and Audio Processing, Vol. 3, No. 1, January 1995.

[10] Sound ideas sound database, http://www.sound-ideas.com