Feature extraction for everyday life sounds

Leila Abdoune[#1], Mohamed Fezari[*2]

[#]*Computer science department, university of Badji Mokhtar Annaba & ENSET*

Skikda, Algeria

[1]lilouchetoday@yahoo.fr, [1]l.abdoune@enset-skikda.dz

[*]*University Badji Mokhtar Annaba, faculty of engineering dept: electronics*

bp:12 Annaba, 23000, Algeria Tel/fax: +21338876565

[2]mourad.fezari@yahoo.fr

Abstract. **Feature extraction is the most crucial step in any process of pattern or sound recognition. The paper considers the task of recognizing environmental sounds for distress situation detection of elderly or disabled and focuses on the study of acoustical parameters of audio signals of everyday life sounds, highlighting the problems related to the definition of the relevant parameters given the variety of environmental sounds and the nature of audio signals. This study is based on a state of the art that shows the various individual solutions but which do not fit with all types of sounds. We try in this paper to show problems and challenges with an approach to address them.**

Keywords— **feature extraction, telemonitoring, everyday life sounds, environmental sound recognition**

I. Introduction

Environmental sound recognition (ESR) is an area that has become increasingly active in recent years. We mean by environmental sounds any sound that can be generated in the environment, natural or artificial, like the sounds of rain, thunder, vehicle, animals, humans, etc. However, there are other definitions for that term. Chachada and Jay Kuo [1] define the environmental sounds like the everyday sounds other than speech and music.

ESR can be used in several applications [1],[15], we mention: robots navigation [12], applications of remote monitoring of elderly and disabled at home [2], [3]. ESR can also be used in smart homes [18]. One of the applications by which we are interested is the remote monitoring of elderly and disabled in their homes by defining first, classes of sounds that can be generated in the restricted environment: home, then trying to recognize the generated sounds to detect a distress situation of the inhabitants due to an incident at home such as the flood or because of a person's fall, etc. After defining the basic sounds or different classes of sounds that can be generated at home in [4] and above that fit with our application which is the remote monitoring of elderly or disabled people, the second step is the feature extraction of sound signals which is a critical step for a recognition or classification system.

In this work, we first present an overview of the final system. After that, a small presentation of everyday life sounds is given in the second section. Thereafter, a description of the problems encountered in the extraction of characteristics of the audio signals is given, followed by a synthesis of some works related to feature extraction of environmental sounds. Finally, concluding remarks and perspectives are given in the last section.

II. Overview of the Final System

Our work is inspired by [13] and [14], which revolve around the environmental sound recognition and classification for surveillance and monitoring applications. Our environment is the habitat of the monitored person. Indeed, in order to recognize the activities of the inhabitants as well as the detection of a possible distress, several sensors can be installed in homes such as cameras, microphones, switches doors, infrared, accelerometer, etc. Each of these sensors can provide a particular type of information: the person's location, its position (lying, standing), the activity carried out, etc. To develop such a system it is important to divide the problem into sub-problems (recognition of activities, recognition of distress...). When all sub systems are developed, we can reach to a complete system that can meet all requirements via the fusion of the resulting data from each system to finally obtain a more accurate decision.

In our work, we are interested in detecting distress situations using the audio channel (microphones installed in the housing). Our system is mainly intended for the recognition of a limited number of classes of environmental sounds, but in a specific environment that is the habitat or home. Environmental sounds with which we are concerned are related to the events of everyday life. Fig. 1 below shows an overview of the overall system and Fig. 2 shows the sound recognition subsystem (system 1).

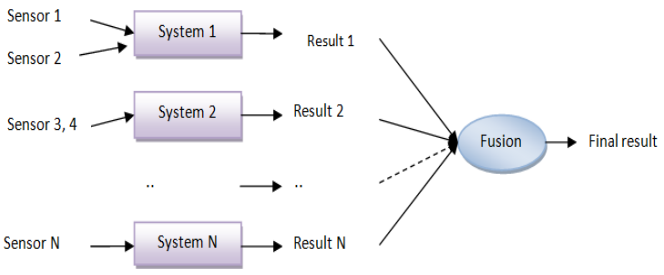

Fig. 1. Global view of the final system

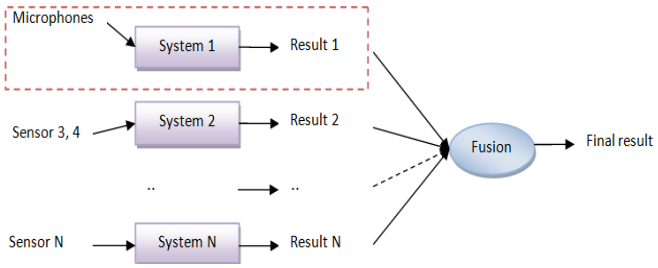

Fig. 2. The sound recognition subsystem

III. Everyday Life Sounds

We mean by everyday life sounds any sound that can be generated in the apartment of the inhabitant namely speech, music, television, dishes sound, slam, opening or closing doors, etc. However, taking into account the nature and the intended application which is remote monitoring of elderly and disabled in their homes, the list of interesting sounds is limited and not all sounds are significant. For this reason, the classes of sounds that seemed interesting were defined in [4], which are divided into four categories, namely: critical sounds, useful sounds, disturbing sounds, and speech (distress keywords). Table 1 below presents the sounds that can be generated in the habitat as well as the necessary sound classes for our application.

TABLE 1

Sounds Generated in the Habitat

Sound			***Speech***	
Critical sounds	***Normal sounds***		***Daily speech***	***Distress words***
	Useful sounds	***Disturbing sounds***		
Screaming, Objects falling, Glass breaking, A long silence.	Dishes, Door opening /closing/ slamming, Footsteps, Water flowing, Yawning,	Television, Radio, Phone ringing, Electrical devices, External noise		I need help, For help, Aie! It hurts, ...

Chachada et al. [1] define the environmental sounds as everyday sounds other than music and speech, therefore everyday life sounds can be environmental sounds, speech and music. If we consider this definition to select the acoustical parameters, we must study feature extraction and selection of audio signals for each of these three types of sounds: speech, music and environmental sounds.

IV. Feature Extraction

A. Feature Extraction Description

A sound recognition application consists of two main modules: a module for extracting the most relevant acoustical features from the input signal, and a pattern recognition module that identifies to which class the signal belongs, which implies the sound identification. Training and classification are not directly made on the acquired signal, but earlier on the features extracted from it. For calculating the acoustical features of a signal, the first step is to make the signal windowing with a specific function. After windowing the signal, features are calculated. Generally, acoustical characteristics can be grouped into two categories: [7] time domain (temporal characteristics) and frequency domain (spectral characteristics). A new taxonomy based on the properties of the audio characteristics was presented in [8].

B. Review of Previous Work

Research in music and speech-like sounds is very advanced compared to environmental sounds. For this reason, several studies have focused on finding the most relevant acoustic parameters to environmental sounds. In this section we will focus on the presentation of these work and essays (table 2).

The work presented in [5], was focused on finding the most effective acoustical features for a system of detection and recognition of sounds for medical supervision. Conventional features such as: signal energy, *Linear predictive coding* (*LPC*), Mel-frequency cepstral coefficients (MFCCs), derived coefficients (Δ, $\Delta\Delta$)) are first tested, and new parameters have been proposed.

In [1], a comprehensive study on feature extraction methods for classification and recognition of environmental sounds, this study was based on the synthesis of work on ESR. Most of these works use the MFCCs, either alone as it gives the best performance or combined with other features to enhance system accuracy. From this study, we can notice that there is not yet a method for the selection of relevant characteristics for ESR applications, this is primarily due to the non-existence of a standard database for the benchmarking of the proposed solutions. Secondly, the compromise between simplicity of the method from time calculation point of view and the effectiveness of the latter. In general, stationary methods are characterized by their simplicity and non-stationary methods are more complex but more effective.

The work presented in [7] deals with the recognition of environmental sounds for the understanding of a scene or the context surrounding an audio sensor. The MP (Matching Pursuit) method was chosen for the selection of the most effective frequency-time domain characteristics, because the use of the frequency domain characteristics only (eg. MFCCs) fails for certain types of sounds and especially noise-like sounds (eg. rain sounds, insect sounds) with a broad flat spectrum. For classification the GMM has been used. To demonstrate the usefulness of the MP-features, tests were made on MFCCs and MP-features and finally, with the combination of MP and MFCCs characteristics. The results of the application are respectively 75.3%, 84.0% and 89.7%. Experimental results show promising performance in the classification 14 different audio environments. The same group in a previous work [20], where the difference is the non-use of the MP characteristics, found the following results for three different classifiers: 96.6% for the SVM, 94.3% for the KNN, and 93.4% for the GMM by using the forward

selection of features. 34 characteristics are used: 1st - 12th MFCCs, Standard Deviation of 1st - 12th MFCCs, Spectral Centroid, Sc, Spectral Bandwidth, Sw, Spectral Asymmetry, Spectral Flatness, Sf, Zero Crossing, Standard Deviation Zero-Crossing, Energy Range, Er, Standard Deviation of Energy Range, Frequency Roll-off, Standard Deviation of Roll-off.

In [9] G. Muhammad and al. propose another method for recognizing environments from audio by combining MFCCs, MPEG-7 descriptors and ZCR (zero-crossing rate) as characteristics. Full use of the MPEG-7 showed improved performance compared to the use of MFCCs. The classifier used is HMM. Experimentation has shown that the combination of these two characteristics gives better performance compared to the use of only MFCCs or MPEG-7 descriptors. When the ZCR is combined with MFCCs and MPEG-7 descriptors, improved performance was observed for some environments.

G. Muhammad and al. in [10] propose a system for the recognition of environments using the MPEG-7 low level audio descriptors with MFCCs. The FDR (fisher Discriminant Ratio) method was used to remove the irrelevant MPEG-7 descriptors, then the PCA (Principal Component Analysis) has been applied to the 30 obtained descriptors to finally get 13 parameters that are combined with MFCCs. The classifier used is the Gaussian mixture model (GMM). The system is evaluated on ten different environmental sounds, the results are encouraging and the proposed system provides significant improvements in the recognition rate. The rate of the proposed recognition system is higher compared to systems based on MFCCs or MPEG-7 descriptors, and this for certain types of environments. In summary, although the MPEG-7 features are more efficient than MFCCs but the combination of the latter two characteristics improves the recognition rate. This work has been also presented in [11] by using MPEG-7 descriptors, temporal ZC (Zero Crossing) and the KNN (k-Nearest Neighbors) classifier.

An approach for classification of locations through the use of audio fingerprint is described in [16]. The number of characteristics is 62 which are time, frequency and statistical domain. Two types of classifiers were used to test the proposed approach: Random Forest and *support vector machines (*SVM). The number of classes is 14 (14 different environments). The results showed that the classification rate is 84.28% for Random Forest and 91.42% for SVMs.

In another experiment, J. Ruben Delgado-Contreras and al. [19] use a feature selection method that is “Chi squared filter” for an application of location classification. The characteristics are then reduced from 62 to 15 (11 statistics and 4 of frequency domain). The classifier used is the SVM, the number of classes is 10 and the recognition rate is higher than 90%.

In the work presented in [6], G.You and al. propose a method called TESPAR (Time Encoded Signal Processing and Recognition) for the recognition of environmental sounds. This method is characterized by its computational resources that are small compared to other methods. To evaluate the proposed system, a comparison was made with a MFCCs based system and a SVM classifier on the same database. The results showed that TESPAR is more effective in the presence of noise and its calculation time is too small compared to the SVM.

The work presented in [17] deals with automatic detection and recognition of impulsive sounds like glass breaking, screams, etc. The system was evaluated on a database which contains 800 signals of 6 different classes. The detection algorithm is based on median filter analyzing the energy changes and it gives good performance even in noise. Two statistical classifiers were used: the GMM and *hidden Markov model* (HMM) to compare results. The results showed that the recognition rate is 98% for an SNR (Signal-to-noise ratio) of 70dB and it is less than 80% for an SNR of 0dB.

Vacher and his co-authors present in [21] AUDITHIS which is a system that performs real-time sound and speech analysis from eight microphone channels. Everyday life sounds are classified with either a GMM or Hidden Markov Model (HMM) classifier. The models were trained with the corpus containing the eight classes of everyday life sounds, using LFCC features (24 filter banks) and 12 Gaussian models. The HMM classifier gives best results in noiseless conditions and the GMM classifier gives best results when the SNR is under +10 dB. Consequently the GMM was chosen as a classifier. The global performance of the system is 72.14% of well-classified sounds.

A. Rabaoui and al. in [22], deal with the supervised classification of sound signals for a monitoring application. The classification method used is one-class SVM. The audio descriptors used are: DWC + MFCC + Energy + Log energy + SRF + SC + ZCR. The rate of good classification is 96.89%. The sounds to recognize are: Emergency Cries, Rifle Shots, Glass Break, Explosions, Door Slam, Dog Barking, Telephone Ringtones, Children's Voices, and Machines.

Another trend in [23], is the proposition of a system for classifying non-speech environmental sounds using a new subset of 2D characteristics, used with a pitch-based (on the pitch range (PR)) feature extraction method and called PR descriptors. Three classifiers are used to evaluate the performance of the system, an SVM (SVM with a linear kernel and SVM with a Gaussian kernel), a neural network with Radial Basis Functions (RBF) and a classifier based on the nearest neighbor method. The sounds to recognize are: Shots, broken glass, cries, dog barking, rain, engine sounds, and restaurant noise. A comparison was made with the same system but using the MFCCs as descriptors. The results of this experiment show that the best recognition rates for the three types of classifiers are obtained by combining the PR and MFCCs descriptors. The results obtained by the MFCCs are better than those obtained by the PRs. For classifiers, SVMs with a linear kernel give an average recognition rate of 85.6% when PR and MFCCs are combined, SVMs with a Gaussian core give a rate of 88.7%, the ANN RBF gives an accuracy of 81.78%, and finally the NN classifier gives a rate of 86.4%. Therefore, in this experiment we find that the SVM with a

Gaussian kernel is the best classifier, and that the PR descriptors combined with the MFCCs give the best results.

From these studies we conclude that most work on environments detection mentioned in the literature use MFCCs, if it is not used alone it is combined with other parameters to improve system accuracy. Characteristics of time-frequency domain are also used for ESR. Other features are also used such as MPEG-7 descriptors and they become more effective when combined with MFCCs. In addition to that, from this comparative study, SVMs are a robust classifiers and generally, give best results in comparison to other classifiers.

TABLE 2

WORK SYNTHESIS

Work	**Objective**	**Audio descriptors**	**Descriptors selection method**	**Classifier**	**Accuracy**
S. Chu and al. [7]	Scene recognition	− MP features, MFCCs	MP (Mutching Pursuit)	GMM	83.9%.
S. Chu and al. [20]	Environment recognition for robots	1st – 12th MFCCs, Standard Deviation of 1st – 12th MFCCs, Spectral Centroid, Spectral Bandwidth, Spectral Asymmetry, .. (32 features)	forward feature selection	-SVM -KNN -GMM	96.6% 94.3% 93.4%
J. R. Delgado-Contreras and al. [16]	Environmental sound recognition	*Temporal* (Short-Time Average Zero-Crossing Rate, ..), *frequency* (Spectral Flux, Spectral Roll Off, Spectral Centroid, Spectral Flatness, Shannon Entropy, ..) and *statistical* (first and second order : Maximum, Minimum, Mean, Median, Standard Deviation, Variance,..) features (62 features)	no	-Random Forest -SVM	84.28% 91.42%
J. R. Delgado-Contreras [19]	Environmental sound recognition	11 statistical et 4 frequency features	Chi squared filter	SVM	higher than 90%
G. Muhammad and al. [9]	Environment recognition	MFCCs, MPEG- 7 (17 temporal and spectral) & ZCR	PCA	HMM	Not shown
G. Muhammad and al. [10]	Environmental sound recognition	MPEG-7 and MFCCs	FDR PCA	GMMs with 4 mixtures	Between 90% 96%
A. Dufaux and al.[17]	Detection and recognition of impulsive sounds	Energy	no	- GMM - HMM	98% at 70dB & above 80% for 0dB
M. Vacher and al. [21]	Recognition of everyday life sounds	LFCC features (24 filter banks)	no	GMM (12 Gaussien models)	72.14%
A.Rabaoui and al. [22]	Environmental sound recognition	DWC + MFCC + Energy + Log energy + SRF + SC + ZCR	Feature vectors selection	1-SVM	96.89%
Uzkent and al. [23]	Environmental sound recognition	PR (pitch range) based features + MFCCs	no	-SVM (gaussien kernel), -ANN (RBF) -KNN	88,7% (Gaussian kernel) 85,6% (linear kernel) 81,78% ANN (RBF) 86,4%. KNN

V. Conclusion

This paper deals with the study of feature extraction of environmental audio signals by analyzing and comparing previous work. Because of the inherent diversity and the unstructured nature of environmental sounds it is difficult to find the characteristics that best describe the audio signals. From this study we concluded that it is difficult to compare the right choice of features in the work done, this is due to the lack of a standard database of everyday life sounds. All the existing works evaluate their systems through the recognition rate, and this assessment is made on their own data (created database and selected sounds). In order to determine the most relevant parameters it is also necessary to take into account the nature of the application; for a real-time application for example, features to choose must not consume an important time in their treatment and the number of features to choose should be restricted to reduce the costs of calculation and execution time but by achieving a good classification rate. Therefore, the appropriate choice of everyday life sounds features is still a very open research area that requires further efforts.

From this comparative study we also conclude that most environments detection works cited in the literature use MFCCs either alone or in combination with other parameters to improve system accuracy. Time-frequency domain characteristics are also used for environmental sound recognition.

Consequently, from this study and from these conclusions we can move to our first experimentations to build a robust system of recognition of everyday life sounds which is the subject of our future work.

References

[1] S. Chachada, and C. C. J. Kuo, "Environmental sound recognition: A survey," 2013 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2013, 2013.

[2] M. Vacher, F. Portet, A.Fleury, and N. Noury, "Challenges in the processing of audio channels for ambient assisted living," in e-Health Networking Applications and Services (Healthcom), 2010 12th IEEE International Conference on. IEEE, pp. 330–337, 2010.

[3] E. Castelli, M. Vacher, D. Istrate, L. Besacier, and J. F. Sérignat, "Habitat telemonitoring system based on the sound surveillance," in: 1st International Conference on Information Communication Technologies in Health , ISBN 960-813-17-1, pp. 141 – 146, Greece, July. 2003.

[4] L. Abdoune and M. FEZARI, "Everyday Life Sounds Database: Telemonitoring of Elderly or Disabled," Journal of Intelligent Systems, ISSN (Online) 2191-026X, ISSN (Print) 0334-1860, 2014.

[5] D. M. Istrate, "Détection et reconnaissance des sons pour la surveillance médicale," Thèse INPG, Spécialité SIPT de l'Ecole Doctorale EEATS, Grenoble, December. 2003.

[6] G. You and Y. Li, "Environmental sounds recognition using tespar. In 5th International Congress on Image and Signal Processing," IEEE, pp. 1796–1800, 2012.

[7] S. Chu, S. Narayanan, and C.-C. J. Kuo, "Environmental sound recognition with time–frequency audio features," IEEE Trans. Audio, Speech, Lang. Process., vol. 17, no. 6, pp. 1142–1158, 2009.

[8] D. Mitrovic, M. Zeppelzauer, and C. Breiteneder,: features for content-based audio retrieval. Advences in computers, vol. 78, pp. 71-150, 2010.

[9] G. Muhammad and K. Alghathbar, "Environment recognition from audio using MPEG-7 features," In: EM-Com proc., pp. 1-6, ISBN: 978-1-4244-4995-8, 2009.

[10] G. Muhammad, Y. A. Alotaibi, M. Alsulaiman, and M. N. Huda, "Environment recognition using selected MPEG-7 audio features and Mel-Frequency Cepstral Coefficients. In Digital Telecommunications (ICDT), 2010 Fifth International Conference on. IEEE, pp. 11–16, 2010.

[11] M. O. AlQahtani, G. Muhammad, and Y. A. Alotaibi, "Environment Sound Recognition using Zero Crossing Features and MPEG-7," In Proceeding of Fifth International Conference on Digital Information Management, Thunder Bay, Ontario, Canada, pp. 502–506, 2010.

[12] N. Yamakawa, T. Takahashi, T. Kitahara, T. Ogata, and H. G. Okuno, "Environmental sound recognition for robot audition using Matching-Pursuit," In Modern Approaches in Applied Intelligence. Springer, pp. 1–10, 2011.

[13] A. Dufaux, "Detection and recognition of impulsive sound signals," PhD thesis, Faculté des sciences de l'Université de Neuchâtel, Suisse, 2001.

[14] M. Cowling, "Non-Speech Environmental Sound Classification System for Autonomous Surveillance," Thesis (PhD Doctorate), Griffith University, Brisbane, 2004.

[15] R.V. Sharan and T.J Moir, "An overview of applications and advancements in automatic sound recognition," Neurocomputing 200: 22-34, 2016.

[16] J.R. Delgado-Contreras, J.P. Garcia-Vazquez, and R.F. Brena," Classification of environmental audio signals using statistical time and frequency features," In: Electronics, Communications and Computers (CONIELECOMP), 2014 International Conference, pp. 212–216 http://dx.doi.org/10.1109/CONIELECOMP. 2014.6808593, 2014.

[17] A. Dufaux, L. Bezacier, M. Ansorge, F. Pellandini, "Automatic sound detection and recognition for noisy environment," In European Signal Processing Conference, Finland, September 2000, pp. 1033–1036, 2000.

[18] J.-C. Wang, H.-P. Lee, J.-F. Wang, and C.-B. Lin, "Robust environmental sound recognition for home automation," IEEE Trans. Autom. Sci. Eng., vol. 5, no. 1, pp. 25–31, 2008.

[19] J. R. Delgado-Contreras, J. P. Garćıa-Vázquez, R. F. Brena, C. E. Galván-Tejada, and J. I. Galván-Tejada, "Feature selection for place classification through environmental sounds," *Procedia Computer Science*, *37*, 40-47, 2014.

[20] S. Chu, S. Narayanan, C. C. J. Kuo, and M. J. Mataric, "Where am I? Scene recognition for mobile robots using audio features," In *Multimedia and Expo, 2006 IEEE International Conference on* (pp. 885-888), IEEE, July. 2006.

[21] M. Vacher, A. Fleury, F. Portet, J.-F. Serignat, and N. Noury, "Complete Sound and Speech Recognition System for Health Smart Homes: Application to the Recognition of Activities of Daily Living," New Developments in Biomedical Engineering, IN-TECH, ISBN-978-953-7619-57-2, pp. 645-67, 2010.

[22] A. Rabaoui, M. Davy, S. Rossignol, Z. Lachiri, and N. Ellouze, "Sélection de descripteurs audio pour la classification des sons environnementaux avec des SVMs mono-classe, " in Actes du 21ème colloque GRETSI : traitement du signal et des images (GRETSI'07), Troyes, France, September. 2007.

[23] B. Uzkent, B. D. Barkana, and H. Cevikalp, "Non-speech environmental sound classification using SVMs with a new set of features," *International Journal of Innovative Computing, Information and Control*, *8*(5), 3511-3524, May. 2012.