A Multidomain Approach for Automatic Home Environmental Sound Classification

Stavros Ntalampiras[1], *Ilyas Potamitis*[2] *and Nikos Fakotakis*[1]

[1] Department of Electrical and Computer Engineering, University of Patras, Rio Patras, 26500 Greece
[2] Department of Music Technology and Acoustics, Technological Educational Institute of Crete, Rethymno, 74100 Greece

sntalampiras@upatras.gr, potamitis@wcl.ee.upatras,gr, fakotaki@upatras.gr

Abstract

This article presents a multidomain approach which addresses the problem of automatic home environmental sound recognition. The proposed system will be part of a human activity monitoring system which will be based on heterogeneous sensors. This work concerns the audio classification component and its primary role is to detect anomalous sound events. We compare the discriminative capabilities of three feature sets (MFCC, MPEG-7 low level descriptors and a novel set based on wavelet packets) with respect to the classification of ten sound classes. These are combined with state of the art generative techniques (GMM and HMM) for estimating the density function of each class. The highest average recognition rate is 95.7% and is achieved by the vector formed by all the feature sets juxtaposed.

Index Terms: computer audition, content-based audio recognition, MPEG-7 audio standard, wavelet packets

1. Introduction

During the last decades there has been an exponential growth of the amount of almost every type of data (audio, video etc) due to the constantly increasing needs in a wide range of applications. The capabilities of personal computers in terms of computational power and capacity have reached very high standards which are provided to the average user with a relatively low price. Additionally, the spread and the raising speed of the global network have made available a great variety of audio data to everyone. In this context, techniques that allow users to navigate through audio data including search and retrieval of sound effects are of key importance and should be designed to operate as efficient as possible. Several requirements have to be met by these algorithms including low computational needs and high accuracy in terms either of classification (recognition rate) or retrieval (false alarm and detection rate).

On top of that, the field of computer audition faces an increasing demand in numerous applications. Humans experience a lot of different types of sounds in their everyday life and have the ability to differentiate them utilizing only the auditory sense. Think as a paradigm the situation where one is sitting inside a living room· using incoming sounds alone one can understand that the radio is on, someone is at the door (doorbell sound) while a baby is crying. The general scope of this work is to build up a system that automatically "understands" its surrounding environment by taking under consideration the sounds it "hears", thus fulfilling the aim for computational auditory scene analysis.

One important application of the generalized sound recognition technology is unattended space monitoring for detecting atypical situations in different kind of environments. Characterization of the surrounding environment in terms of threatening/non-threatening situation would be very useful towards preventing loss of human life and/or property damage. A detailed survey of the specific type of frameworks can be found in [1]. Additionally, there are many other real-world applications of this technology such as environmental monitoring, music processing and bioacoustic identification [2-4].

The closest paper to this work that essentially uses audio to identify sounds that exist inside a typical home environment is [5]. They employ low level descriptors from MPEG-7 audio protocol in combination with hidden Markov models to reach 95% average recognition rate while seven sound classes were considered (cat, dog, male and female speech, doorbell, noc and laugh). Their corpus consisted of 500 sounds taken from various sources including the TIMIT speech database and Sounddogs sound effects library. In this paper we propose a probabilistic framework for automatic recognition of home environmental sound events. Ten sound categories are organized using several databases and features derived from frequency and wavelet domains are extracted. Descriptors which belong to different domains reflect better upon diverse aspects of the audio structure. Even though they sometimes offer redundant information and increase the computational complexity, they could provide improved performance under adverse conditions where portions or certain frequency bands of the signal are corrupted or absent. In this paper we use three different feature sets separately while their simultaneous usage is shown to produce the best average classification accuracy.

The novel aspects of the presented approach are the investigation of features which belong to different domains for classifying a wide spectrum of home environmental sounds, including abnormal sound events and applying different pdf estimation algorithms for the needs of the specific application. Furthermore we organized a thorough and concise corpus which consists of professional sound effect collections. We also considered a class with baby crying sound events. Detection of such events can be of great importance for alerting parents, thus helping them to handle the particular situation [6]. It should be mentioned that this work is only an initial step towards achieving the goals of PROMETHEUS project (http://www.prometheus-fp7.eu) which intends to construct a synergetic network of heterogeneous modalities. The different types of information will be fused though a probabilistic framework which will serve understanding of human behavior as well as complex human interaction.

The organization of this work is the following: in the next section we describe all the involved feature sets. Sections 3 and 4 present the pattern recognition module and the experimental protocol respectively. The last section includes our conclusions and future work.

2. Acoustic Parameters

In this section we analyze the group of descriptors that were used for training statistical models the parameters of which represent the *a-priori* knowledge we have about the distribution of the sound categories. We selected Mel scale filterbank because of its ability to capture the most important information as regards human perception. Furthermore the MPEG-7 standard currently constitutes the state of the art methodology for unsupervised content based generalized audio recognition while the third set is based on multiresolution analysis – perceptual wavelet packets – which still haven't gained much attention regarding to the non-speech audio processing field. The parameters (frame, overlap, FFT size etc) that were used during the extraction process of all sets were exactly the same so as to have a reliable comparison of the performance that each one achieved.

Silence is considered to be "noise" in this particular task making harder the modeling process, hence reducing the probability of correct classification. Thus silence was removed using a statistical model-based voice activity detector [7] before feature extraction. Furthermore mean value removal was applied onto the sampled waveforms to eliminate any possible DC-offset while no pre-emphasis was employed. The next paragraphs describe the processes that lead to each set's extraction from the silence free signals.

2.1. Mel-Frequency Cepstral Coefficients (MFCC)

This feature set is composed of the first thirteen Mel frequency cepstral coefficients including the 0^{th} coefficient which reflects upon the energy of each frame. For MFCC's derivation we compute the power of the short time Fourier transform with respect to every frame and pass them through a triangular Mel scale filterbank, which emphasizes signal components which play an important role to human perception. Subsequently, the log operator is applied and the energy compaction properties of discrete cosine transform are exploited in order to decorrelate and represent the majority of each band-energy with just a few coefficients. Lastly a thirteen-dimension vector is formed by the most important thirteen coefficients. It should be mentioned that cepstral mean normalization was applied.

2.2. MPEG-7 Low Level Descriptors (LLDs)

The respective audio protocol offers a variety of standardized tools for automated audio content description in terms of both low and high level meaning extraction. Not only LLDs are provided but a degree of "explanation" of a specific audio waveform using Description Schemes, which try to bridge the gap between the low-level features and the semantic level that is desirable for interpretation and understanding of the soundscene. After extensive experimentations as regards all the parameters provided by the standard, in this work we concluded to use the LLDs tabulated in Table I. A short description of the selected features can be found in previous work of ours while a detailed commentary in [8].

Descriptor	Dimensions	Abbreviation
Audio Waveform	2	AW
Audio Power	1	AP
Audio Spectrum Centroid	1	ASC
Audio Spectrum Spread	1	ASS
Audio Spectrum Flatness	19	ASF
Harmonic ratio	1	HR
Upper Limit of Harmonicity	1	ULH
Audio Fundamental Frequency	1	AFF

Table 1. *MPEG-7 LLDs descriptors and dimensions.*

2.3. Perceptual Wavelet Packet Analysis

Regarding the third feature set, we introduce the usage of critical band-based multiresolution analysis for automated home environmental sound classification. Lately, digital signal processing using wavelets has become a common tool in many diverse research areas. Some examples are bioacoustic signal enhancement [9] and applications in geophysics (tropical convection, the dispersion of ocean waves etc) [10]. The main advantage of the wavelet transform is that it can process time series, which include non stationary power at many different frequencies. The fundamental property of the Fourier transform is the usage of sinusoids with infinite duration. While they are smooth and predictable, wavelets tend to be irregular and asymmetric. They comprise a dynamic windowing technique which can treat with different precision low and high frequency information. The first step of the wavelet packet analysis is the choice of the original (or mother) wavelet and by utilizing this function, the transformation breaks up the signal into shifted and scaled versions of it. In this paper we utilized Daubechies 1 (or Haar) function as the original wavelet. Unlike discrete wavelet transform (DWT), when wavelet packets (WP) are employed both low and high frequencies coefficients are kept. In our case the DWT is applied three subsequent times and consists of three-stage filtering of the audio signals as we can see in Fig. 1.

The idea behind the third set is the production of a vector that provides a complete analysis of the audio signal across different spectral areas while they are approximated by WP. We should also take into account that not all the parts of the spectrum affects human perception the same (which is crucial for sound recognition), thus the division of the spectrum requires a fine partioning. In [11], it is observed that the human auditory system filters the entire audible spectrum into many critical bands. Based on this observation, we employed a critical-band based filterbank using Gabor bandpass filters. Subsequently three-level wavelet packets are extracted out of each spectral band. Downsampling is applied on each coefficient at each stage in order not to end up having the double amount of data, as Nyquist theorem requests. The wavelet coefficients are then segmented and the autocorrelation envelope area is computed and normalized by half the segment size. N normalized integration parameters are calculated for each frame, where N is the total number of the frequency bands multiplied by the number of the wavelet coefficients (17x8=136). This series of parameters comprises the PWP-Integration feature vector and the entire calculation process is depicted in Fig. 1 (a ready to run implementation of

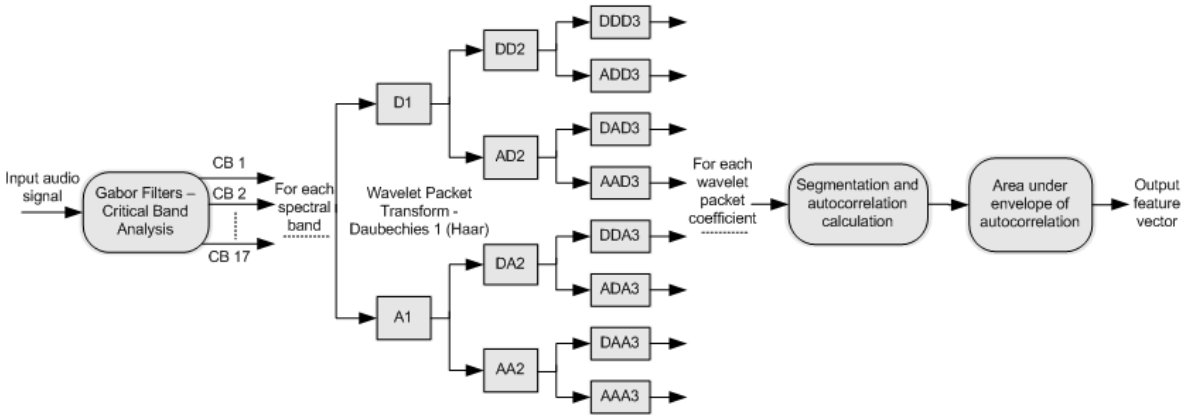

Figure 1: *Perceptual wavelet packet integration audio analysis.*

this feature extraction methodology is available at http://www.wcl.ece.upatras.gr/dalas).

3. Pattern Recognition Schemas

We compared the recognition performance of two generative approaches which are based on the underlying assumption that the data belonging to a specific class are described probabilistically by a mixture of Gaussian distributions. The main characteristic of this type of classifiers is that they handle the samples of each class independently of the other classes. We employed Gaussian mixture models and Hidden Markov models with two different topologies (left-right and fully-connected). Unlike GMMs, HMMs have the ability to model the temporal evolution of sound events. This kind of pdf approximation is based on the assumption that the data follow a finite temporal pattern which is expressed in terms of states. Subsequently the previously created models are used for computing a degree of resemblance (e.g. log-likelihood) between each model and an unknown input signal. This type of score is compared against the rest and the final decision is made with a maximum log-likelihood determination. Torch (available at http://www.torch.ch) implementation of GMM and HMM, written in C++ was used during the whole process. The maximum number of k-means iterations for initialization was 50 while both the EM had and upper limit of 25 iterations with a threshold of 0.001 between subsequent iterations.

4. Experimental Set-Up

In general, the specific problem concerns numerous sound sources which are often encountered in a standard home environment. Many of these sources are unpredictable while the most frequent ones are: speech, music, door knock, dog bark, cat meow, door dell and laughter while the group of atypical sound categories includes baby crying, gunshot and explosion. Due to the unavailability of such a corpus, a combination of databases was necessary. The following datasets were employed: (i) BBC Sound Effects Library, (ii) Sound Ideas Series 6000, (iii) TIMIT, (iv) Sony Sound Effects Library, (v) Sound Ideas: the art of Foley and (vi) an EBU music collection. The final audio classes can be observed in Table 2. These audio recordings are of high quality while they are widely used by the movie industry. It is usually the case that the audio stream of a movie is processed or even replaced entirely. Thus there exist massive data collections for the construction of trained probabilistic models for audio event detection and classification. The concurrent usage of these datasets offers great variability and diversity regarding the *a-priori* knowledge which is to be incorporated to the probabilistic models. As long as the audio samples include all the possible realizations of the sound classes, it is safe to assume that the constructed models represent the specific categories appropriately. The dataset was randomly divided into 70% for model training and 30% for testing. All the audio sequences were downsampled to the sampling rate of 16 KHz with 16bit analysis which is appropriate for recognition purposes.

The number of Gaussian components was varying from 2-64 with step 1 while the number of states was taken from the set {3, 4, 5, 6, 7}. The final choice was made after extensive experimentations using the highest average recognition rate criterion. The performance was measured using per frame audio analysis on novel data.

The average recognition rates with respect to each feature set are tabulated in Table 3. The best performance is provided for each type of classifier along with the corresponding parameters. As we can see the feature set with the best performance is the one based on MPEG-7 audio protocol LLDs (90.7%) while using ergodic HMMs. The second best performance is achieved by the group based on wavelet packets (87.1%) combined with the same type of classifier. MFCCs demonstrated the worst performance (86.9%) while left-right HMMs were employed for class modeling. In general we can say that for some classes the different sound parameters exhibit similar results. For example they classify with high accuracy speech, music and laughter sound events.

Audio Category	**No. of Sound Samples**	**Total Duration (sec)**
Male and female speech	1680	5.073,6
Music	54	1.209,6
Door knock	287	1.004,5
Dog barking	102	1.101,6
Cat meowing	143	886,6
Baby crying	67	2.546
Door bell	118	1.368,8
Laughter	145	3.161
Gunshot	131	1.803,87
Explosion	187	6.159,78
Total	**2.914**	**24.315,35**

Table 2. *Audio classes of the final dataset.*

Feature set	Type of classifier	No. modes	No. states	Average recognition rate
MFCC	GMM	9	-	82.1
	Left-right HMM	13	3	**86.9**
	Ergodic HMM	8	3	83.3
MPEG-7	GMM	20	-	85.5
	Left-right HMM	10	4	86.4
	Ergodic HMM	11	5	**90.7**
PWP Integration	GMM	16	-	80.6
	Left-right HMM	16	3	82.9
	Ergodic HMM	16	3	**87.1**
Combination	GMM	8	-	87.9
	Left-right HMM	32	3	**95.7**
	Ergodic HMM	16	6	92.3

Table 3. Average r*ecognition rates (%) with respect to each feature set and classifier type.*

However there are some classes where the results differ· such a case is the door bell class. Therefore different sound classes are categorized by different feature sets with different accuracies. Following this observation we decided to juxtapose all three feature sets and use them combined for model training and testing so as to exploit their complementary character as regards several partial classification tasks. This experiment gave out the highest average recognition rate (95.7%) while left-right HMMs were used. We observed that all the partial classification rates are increased since the feature sets capture diverse aspects of the audio signal structure. Several or the errors occur due to the high within-class variability while some audio samples are acoustically similar even though they belong to different sound categories. We conclude that the results are more than promising and underline the importance of the probabilistic structure constructed using feature sets of diverse domains.

5. Conclusions

In this paper we thoroughly investigated the performance of three groups of sound parameters for home environmental sound recognition through a probabilistic framework. A feature set based on wavelet packets was introduced and it was shown that it achieves better performance than the MFCC set. The vector which is comprised by the coefficients of the three feature sets together, reached the highest average classification rate. We intent to enhance the performance of the presented acoustic monitoring component by incorporating other modalities and perform audio event detection using non-audio capturing devices too, e.g. a person standing at the door raises the probability of a door bell sound event. We wish to fuse the heterogeneous modalities on the feature, probability and decision level. Furthermore this component will contribute to high-level scene analysis by recognition of human activities in a home environment.

6. Acknowledgements

This work was supported by the EC FP 7th Grant PROMETHEUS 214901 “Prediction and Interpretation of human behaviour based on probabilistic models and heterogeneous sensors.”

7. References

[1] Ntalampiras, S., Potamitis, I. and Fakotakis, N., "An adaptive framework for acoustic monitoring of potential hazards", EURASIP Journal on Audio, Speech and Music Processing, doi:10.1155/2009/594103, http://www.hindawi.com/journals/asmp/2009/594103.html

[2] Kwan, C. , Ho, K. C., Mei, G., Li, Y., Ren, Z., Zhang, Y., Lao, Stevenson, M., Stanford V. and Rochet, C., "An automated acoustic system to monitor and classify birds", EURASIP Journal on Applied Signal Processing, pp. 1-19, 2006.

[3] Casey, M., Veltkamp, R., Goto, M., Leman, M., Rhodes C. and Slaney, M., "Content-based music information retrieval: current directions and future challenges", Proceedings of the IEEE, 96(4):668-696, 2008.

[4] Lee, C.-H., Han, C.-C. and Chuang, C.-C., "Automatic classification of bird species from their sounds using two-dimensional cepstral coefficients", IEEE Transactions on Audio, Speech and Language Processing, 16(8):1541-1550, 2008.

[5] Wang, J.-F., Wang, J.-C., Huang, T.-H. and Hsu, C.-S., "Home environmental sound recognition using MPEG-7 features", in IEEE International Symposium on Micro-Nano Mechatronics and Human Science, December 2003.

[6] Ruvolo, P. and Movellan, J., "Automatic cry detection in early childhood education settings", MPLab Technical Report 2008-3.

[7] Sohn, J., Kim, N. S. and Sung, W., "A statistical model-based voice activity detection", in IEEE Signal Processing Letters, 6(1):1-3, 1999.

[8] Kim, H.-G., Moreau, N. and Sikora, T., MPEG-7 Audio and Beyond: audio content indexing and retrieval, Wiley Publishers, October 2005.

[9] Ren, Y., Johnson, M. T. and Tao, J., "Perceptually motivated wavelet packet transform for bioacoustic signal enhancement", Journal of Acoustic Society of America, 124(1):316-327, 2008.

[10] Torrence, C. and Compo, G. P., "A practical guide to wavelet analysis", Bulletin of the American Meteorological society, 79(1):61-78, 1998.

[11] Yost, W. A., Fundamentals of Hearing, 3rd Edition, New York Academic, pp. 153-167, 1994.