31st Annual International Conference of the IEEE EMBS
Minneapolis, Minnesota, USA, September 2-6, 2009

Audio Sound Event Identification for Distress Situations and Context Awareness

J.E. Rougui, D. Istrate and W. Souidene

Abstract— **In this paper the acoustic event detection and classification (AED/AEC) system developed under European Community's Seventh Framework Companionable project in awareness context is presented. The system relies on the use of Wavelet transform technique for detection and on an unsupervised order estimation of Gaussian mixture model (GMM) arranged in hierarchical form in the aim to improve the recognition accuracy. The results, measured in terms of two metrics (accuracy and error rate) are obtained applying the implemented system in off-line mode of audio analysis form of distress scenarios recorded in this fact.**

I. INTRODUCTION

The information from the everyday life sound flow is more and more used in telemedical applications in order to detect falls, to detect daily life activities or to characterize physical status. The use of sound like an information vector has the advantage of a simple and cheapest sensors, is not intrusive and can be fixed in the room. Otherwise, the sound signal has important redundancy and need specific methods in order to extract information. The definition of *signal* and *noise* is specific for each application; e.g. for speech recognition, all sounds are considered noise. Between numerous sound information extraction applications we have the characterization of cardiac sounds [1] in order to detect cardiac diseases or the snoring sounds [2] for the sleep apnea identification.

Age of population increases in all societies and the elderly people prefer to preserve their independence at home. The remote monitoring, a telemedicine application, represent a real solution to the lack of medical staff and to ensure the safety at home. At home, for the elderly people the falls and faintness are the most important factor of accidents. Many research laboratories has proposed the fall detection using wearable devices[3], [4] but the person need to carry the device permanently. Non intrusive solution, using the analysis of activity daily life (ADL) using behavioral patterns [5] or an anxiety model [6] has been proposed . The detection of ADL can be carry out trough electrical sensor, door sensors, infrared, wearable or sound [7].

Using sound for the fall detection has the advantage that the patient doesn't have to carry a wearable device but less robust in the noise presence and depend from acoustic conditions [8], [9]. The combination of several modalities in order to detect distress situation is more robust using the information redundancy. We have proposed a first approach of data fusion between a wearable device, infrared sensors and sound analysis [10].

We have already proposed a sound environment analysis system for remote monitoring [11], capable to identify everyday life normal or abnormal and distress expressions. Currently in the framework of the CompanionAble project[1] we participate in developing a coupled smart sensor system with a robot for mild cognitive impairment patients. The sound modality is used like a simplified patient-system interface and for the distress situation identification. The sound system will participate to the context awareness identification, to the domotic vocal commands and to the distress expressions/sounds recognition.

II. AUDIO SOUND EVENT IDENTIFICATION

Currently, the ability to identify sounds when other noises are present remains a difficult task as far as the speech features are designed and mainly based on the properties of speech production and perception. Moreover, to simulate the non-uniform frequency resolution observed in human auditory perception, these speech features set adopt uniform critical bands, providing high resolution in a low frequency part. However, the spectral structure of acoustic events is different from that of speech [12].

Alternatively, to settle partially with this problem, the sound event identification system proposed in this work, deals with the use of the unsupervised estimation of the variability spectral of each acoustic event detected. The classification module exploits the efficiency representation of Mel Frequency Cepstral Coeficient (MFCC) features extracted from the whole 24 Mel filter using EM and another criterion was suggested by Rissanen [13] called the minimum description length (MDL) estimator. This estimator works by attempting to find the model order which minimizes the number of sub clusters by combining two nearest clusters.

The proposed audio sound event identification system is composed by several modules (Fig. 1.):

- Sound event segmentation enable detect the sound events presence or change,
- Sound/Speech classification,
- Sound classification identifying the reference of sound,
- Noise classification which can be superposed to an another sound.

This work was supported by European Community's Seventh Framework Programme (FP7/2007-2011), CompanionAble Project (grant agreement n. 216487).

J.E. Rougui, D. Istrate and W. Souidene are with LRIT-ESIGETEL, Rue du Port de Valvins, 77210 Avon, France {firstname.lastname}@esigetel.fr

[1] www.companionable.net

978-1-4244-3296-7/09/$25.00 ©2009 IEEE

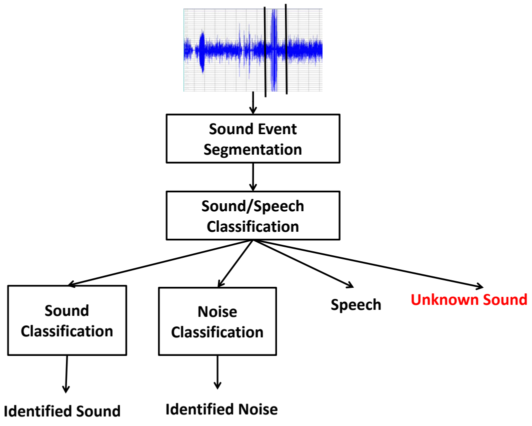

Fig. 1. Audio Event Identification architecture

The structure shown above is hierarchical form, the goal is to carried out successive tests in order to refine the event alarming retrieval. The technique of test rejection is used in each test level. The proposed algorithm is robust even when the tested conditions are completely different from the conditions where the training signals were recorded. The query utterance segment is labeled ”Unknown” in three cases:

- the query segment for test is insufficient,
- the query segment is identified by a ”Unknown” model.

A. Segmentation Module

The general scheme of the proposed system of event segmentation with adaptive threshold is shown in Fig.2. Firstly, the initialization of threshold parameters is arbitrary. Furthermore, based on this technique no pre-processing is required. Compared with other segmentation techniques like BIC-based segmentation, which depends on the observation window adjustement and a simple Gaussian representation, the proposed segmentation is based on the Discret Wavelet Transformation (DWT) coefficients in the high order bands in order to detect the sound events presence or change continuously, a constraint imposed by the real-time processing. The DWT windows is 128 ms the algorithm of *flowchart*[14], calculates the energy of the 8, 9 and 10 represent the high order of wavelet coefficients (frequencies more than 2kHz). Then, the DWT energy value is compared with an adaptative threshold presented in (Eq.1).

$$E\_th = detect\_threshold + Overlay * Average\_E \quad (1)$$

The algorithm of adaptive threshold used for event segmentation, benefit of the effectiveness of the distribution of the energies among blocks of wavelet packet coefficients used to detect the change of spectral information presented in high frequency. Therefore, this algorithm is extended to be used in difficult condition of acquisition using a omnidirectional microphone or a several microphones in facilities suitable for Smart-Home. The fixed part of the threshold ($detect\_threshold$) became self adapted to deal with the intra-variability of the energy amplitude of the same event recorded in different distance and the extra-variability due of the number of microphones capture and localization. Firstly, as the Fig.2 shown, the threshold adaptation is occurred in three cases:

1) after the event detection, the threshold is adapted according to the energy mean observed over some windows after the end of event;
2) the detection of the increasing value of energy between two observation window, i.e, the peak detection technique is applied using the curve given by the energy calculated from the wavelet coefficients.

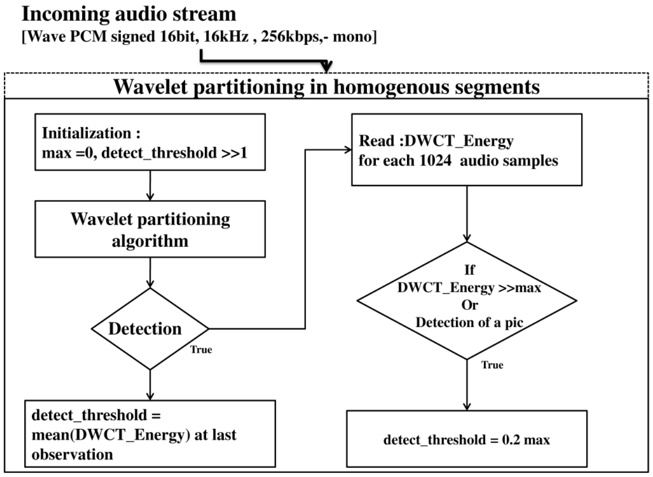
Incoming audio stream
[Wave PCM signed 16bit, 16kHz , 256kbps,- mono]
Wavelet partitioning in homogenous segments
Initialization : max =0, detect_threshold >>1
Wavelet partitioning algorithm
Detection
True
detect_threshold = mean(DWCT_Energy) at last observation
Read :DWCT_Energy for each 1024 audio samples
If DWCT_Energy >>max Or Detection of a pic
True
detect_threshold = 0.2 max

Fig. 2. Threshold adaptation

B. Hierarchical classification module

1) Audio features employed: The sound signal is sampled at 16kHz, and framed length/shift is 16ms/8ms using Hamming window. For each frame, a set of Cepstral coefficients has been extracted using a Fourier Transform and to map the powers of the spectrum obtained above onto the Mel scale, using triangular overlapping windows. In this work the MFCC value is used in a classical form in order to test the identification system of sound events when other noises are present in Smart-Home environment. Recall, the MFCC parameters are increasingly finding uses in music information retrieval applications such as genre classification, audio similarity measures, and recently these have been successfully used in audio-based context recognition [15], [16].

2) Proposed technique : unsupervised sound model representation: The present work is set in the framework of probabilistic modeling and statistical decision criteria, which is common and effective for sound recognition. A classical solution to the above mentioned task would consist in partitioning the audio sources in events-homogeneous segments, then the segments is classified with statistical model using Gaussian Mixture Model (GMM).

The technique disclosed here exploits the fact the features extracted for vocal or non-vocal event using 24 MFCC coefficients features take the form of a Gaussian mixture model with full covariance matrices. With this form, the data related to event k, i.e. mel-cepstral feature vectors, is assumed to be drawn from a probability density.

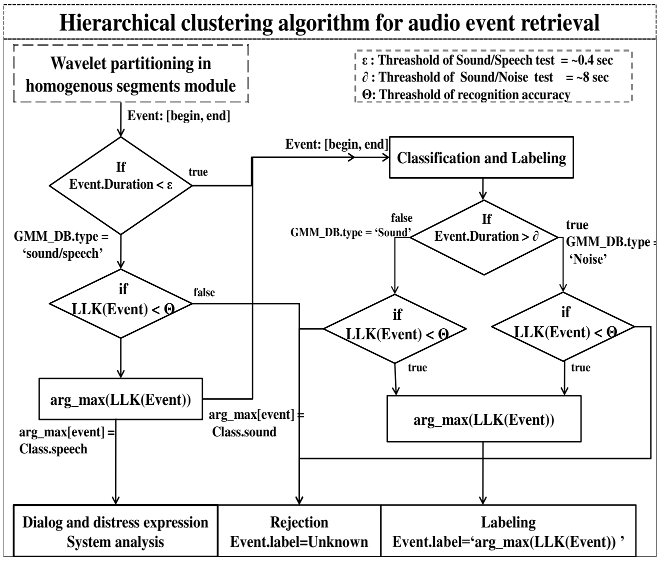
Hierarchical clustering algorithm for audio event retrieval
Wavelet partitioning in homogenous segments module
ε : Threashold of Sound/Speech test = ~0.4 sec
∂ : Threashold of Sound/Noise test = ~8 sec
Θ: Threashold of recognition accuracy
Event: [begin, end]
If Event.Duration < ε
true
GMM_DB.type = 'sound/speech'
if LLK(Event) < Θ
false
arg_max(LLK(Event))
arg_max[event] = Class.speech
arg_max[event] = Class.sound
Event: [begin, end]
Classification and Labeling
false
GMM_DB.type = 'Sound'
If Event.Duration > ∂
true
GMM_DB.type = 'Noise'
if LLK(Event) < Θ
true
if LLK(Event) < Θ
true
arg_max(LLK(Event))
Dialog and distress expression System analysis
Rejection Event.label=Unknown
Labeling Event.label='arg_max(LLK(Event)) '

Fig. 3. Hierarchical sound retrieval strategy

It is often desirable to model distributions that are composed of distinct subclasses of clusters. For example, the event spectral information might behave differently according to the power spectrum for the fluctuations caused by the vibrations, sound propagation etc. Therefore, the aggregate behavior is likely to be a mixture of a different and distinct behavior. The main role of the mixture distributions is to form a probabilistic model composed of a number of component subclasses. Each subclass is than characterized by a set of parameters describing the mean and the variance of the cepstral components.

In the training stage, the sound/speech/noise signals represented by the corresponding distribution of features are approximates with a Gaussian mixture density function. The proposed technique, provide for each sound type a representative GMM model with a different order according to the spectral signature of the event corresponding and the training data duration (see Table I). Moreover, The sound event retrieval technique is shown in Fig.3 based on Hierarchical clustering algorithm.

III. Sound Corpus

The sound classification module has currently 11 sound classes trained on 70 minutes of signal and 5 noises of everyday life (Vacuum cleaner, Water flushing, Dishwasher, Hairdryer) trained on 18 minutes of signal (Table I). The classes referring to sound and speech for the first classification level were trained on all existing sounds and on 38 minutes of speech respectively.

IV. Evaluation

The acoustic event detection and classification proposed system was evaluated in terms of sound recognition in the noise presence and globally on a generated scenario.

A. Noise influence on sound recognition

The ten sound classes trained on 90% of the pure sounds was used to evaluate the error recognition rate in the case of adding different noises at 4 signal to noise ratio (SNR): pure (70 dB), 40 dB, 20 dB and 10 dB. The analyzed noises was specific everydaylife noises: electric shaver, hairdryer, water flushing and vacuum cleaner. This type of noise in the most of the cases can catch an useful sound like glass breaking or object fall which is crucial to detect a distress situation. Analyzing the results from Table II we can observe that the sound recognition has good performances for the electric shaver and for hairdryer because of its spectral composition. The worst results are obtained in the presence of vacuum cleaner and water flushing at 0 dB of SNR. Our futur researches will be focused more on the reducing the influence of these two type of noise taking into account that an important part of distress situation take place in the bathroom where the only sensors are the sound and the infrared one.

TABLE I
Data Base of Sounds

Sound Event	File numbers	Duration	GMM Order
Door	523	358 s	22
Glass breaking	88	75 s	4
Phone Ringing	517	75 s	28
Key ringing	22	469 s	22
Door Bell	17	53 s	4
Cry-scream	73	29 s	2
Dishes	163	65 s	27
Yawn	9	87 s	25
Object Falls: metal	2	660 s	23
Paper	21	760 s	20
Chair	10	727 s	24
Speech	2646	2321 s	15
Sound	1445	2347 s	29
Vacuum cleaner	8	234 s	22
Water flushing	2	35 s	4
Dishwasher	47	436 s	9
Hairdryer	34	175 s	8

TABLE II
Sound error recognition rate in the noises presence

Type of noise	SNR			
	70 dB	40 dB	20 dB	10 dB
Electric Shaver	0.83 %	0 %	2.48 %	7.44 %
Hairdryer	0.83 %	0 %	2.48 %	8.26 %
Water flushing	0.83 %	1.65 %	11.57 %	34.71 %
Vacuum cleaner	0.83 %	0.83 %	14.88 %	51.24 %

B. Global system evaluation on audio scene scenario

The Global acoustic event detection system was evaluated on a generated scenarios containing 9 types of sounds and speech mixed with a 5 different noises at SNR varying between 10 and 40 dB. The time between two acoustic events is randomly chosen between 1 and 3 seconds. The scenario duration is about 164 seconds and is contains 50 acoustic events.

The sound event detection is evaluated in terms of number of correct detected events. We consider a correctly detected event if the middle of segmented signal corresponds to a reference segment and if its dimension is at minimum about 50% of the reference one. The Acoustic Event Detection rate (AED) is computed using 2.

$$AED = \frac{\#\, correct\ detected\ events}{\#\, events} * 100 \tag{2}$$

The classification sound/speech and sound classification are evaluated in terms of correctly classified signals through Acoustic Event Classification (AEC) using 3.

$$AEC = \frac{\#\, correct\ classified\ detected\ events}{\#\, correct\ detected\ events} * 100 \tag{3}$$

The global performances of AED system are evaluated trough Acoustic Events Detected and Classified Rate (AEDC) using 4.

$$AEDC = \frac{\#\, correct\ classified\ events}{\#\, events} * 100 \tag{4}$$

The results presented in the Table III confirm the good detection rate of the sound segmentation module (AED) and also the recognition rate in the noise presence (AEC). Globally the proposed system has an error recognition rate about 25% when the SNR is greater than or equal to 20db.

V. CONCLUSIONS AND FUTURE WORKS

In this paper we present an acoustic event detection and classification system which is based on an already proposed one with the optimization of the sound segmentation module and also of the sound recognition one. This system belongs to CompanionAble project with the aim to detect sounds, distress expressions and vocal commands. The current performances need to be ameliorated in the presence of water flushing and other bathroom noises. Future work could consist in adopting of other descriptors. Furthermore, likelihood based score normalization will be investigated in order to compare between GMM with different components order.

Currently, we study the coupling between proposed system and a CMT microphone in order to localize and enhance the sound signal (collaboration with AKG partner).

TABLE III

GLOBAL SYSTEM PERFORMANCES IN THE NOISE PRESENCE

Performances	SNR			
	Noise	40 dB	20 dB	10 db
AED	Apartment noise	97.96 %	87.96 %	71.43 %
	HairDrayer	83.76 %	32.65 %	22.45 %
	Vacuum cleaner	97.96 %	87.76 %	71.43 %
	Electric Shaver	81.63 %	30.61 %	22.45 %
	Water flushing	85.71 %	36.73 %	14.29 %
AEC	Apartment noise	89.58 %	90.70 %	80.00 %
	Hairdrayer	90.24 %	68.75 %	81.82 %
	Vacuum cleaner	89.19 %	80.00 %	54.55 %
	Electric Shaver	90.00 %	73.33 %	81.82 %
	Water flushing	90.48 %	77.78 %	57.14 %
AEDC	Apartment noise	87.76 %	79.59 %	57.14 %
	HairDrayer	75.51 %	22.45 %	18.37 %
	Vacuum cleaner	76.74 %	40.00 %	18.75 %
	Electric Shaver	73.47 %	22.45 %	18.37 %
	Water flushing	77.55 %	57.14 %	8.16 %

ACKNOLEDGEMENTS

The authors gratefully acknowledge the contribution of European Community's Seventh Framework Programme (FP7/2007-2011), CompanionAble Project (grant agreement n. 216487).

REFERENCES

[1] C. S. Lima and D. Barbosa, "Automatic segmentation of the second cardiac sound by using wavelets and hidden markov models," in *IEEE EMBC*, Vancouver, Canada, Aug 20-24 2008, pp. 334–337.

[2] A. K. Ng and T. S. Koh, "Using psychoacoustics of snoring sounds to screen for obstructive apnea," in *IEEE EMBC*, Vancouver, Canada, Aug 20-24 2008, pp. 1647–1650.

[3] E. Campo and E. Grangereau, "Wireless fall sensor with gps location for monitoring the elderly," in *IEEE EMBC*, Vancouver, Canada, Aug 20-24 2008, pp. 498–501.

[4] N. Noury, A. Tarmizi, *et al.*, "A smart sensor for the fall detection in daily routine," in *SICICA 2003*, Aveiro, Portugal, July 2003.

[5] G. Virone, M. Alwan, S. Dalal, S. W. Kell, B. Turner, J. A. Stankovic, and R. Felder, "Behavioral patterns of older adults in assisted living," *IEEE Transactions on TITB*, vol. 12, no. 3, pp. 387–398, 2008.

[6] S. Moncrieff, S. Venkatesh, G. West, and S. Greenhill, "Incorporating contextual audio for an actively anxious smart home," in *Intelligent Sensors, Sensor Networks and Information Processing Conference*, December 2005, pp. 373–378.

[7] D. Maunder, E. Ambikairajah, J. Epps, and B. Celler, "Dual-microphone sounds of daily life classification for telemonitoring in a noisy environment," in *IEEE EMBC*, Vancouver, Canada, Aug 20-24 2008, pp. 4636–4639.

[8] M. Popescu, Y. Li, M. Skubic, and M. Rantz, "An acoustic fall detector system that uses sound height information to reduce the false alarm rate," in *IEEE EMBC*, Vancouver, Canada, Aug 20-24 2008, pp. 4628–4631.

[9] D. Litvak, Y. Zigel, and I. Gannot, "Fall detection of elderly through floor vibrations and sound," in *IEEE EMBC*, Vancouver, Canada, Aug 20-24 2008, pp. 4632–4635.

[10] W. Souidene, D. Istrate, H. Medjahed, J. Boudy, J. L. Baldinger, I. Belfeki, and J. F. Delavaut, "Multi-modal platform for in-home healthcare monitoring (emutem)," in *International Conference on Health Informatics (HEALTHINF 2009)*, Porto, Portugal, January 2009, pp. 381–386.

[11] D. Istrate, M. Binet, and S. Cheng, "Real time sound analysis for medical remote monitoring," in *IEEE EMBC*, Vancouver, Canada, Aug 20-24 2008, pp. 4640–4643.

[12] X. Zhou, X. Zhuang, M. Liu, H. Tang, M. Hasegawa-Johnson, and T. S. Huang, "Hmm-based acoustic event detection with adaboost feature selection," in *CLEAR*, 2007, pp. 345–353.

[13] J. Rissanen, "A universal prior for integers and estimation by minimum description length," in *Ann. Statist.*, vol. 11, 1983, pp. 416–431.

[14] D. Istrate, E. Castelli, M. Vacher, L. Besacier, and J. Serignat, "Information extraction from sound for medical telemonitoring," *IEEE Transactions on TITB*, vol. 10, pp. 264–274, April 2006.

[15] A. Eronen, J. Tuomi, A. Klapuri, D. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio–based context awareness – acoustic modeling and perceptual evaluation," 2003.

[16] A. Eronen, V. T. Peltonen, J. T. Tuomi, A. Klapuri, S. Fagerlund, T. Sorsa, G. Lorho, and J. Huopaniemi, "Audio-based context recognition." *IEEE Transactions on Audio, Speech & Language Processing*, no. 1, pp. 321–329.