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

Audio Based Surveillance for Cognitive Assistance Using a CMT Microphone within Socially Assistive Technology

J.E. Rougui, D. Istrate, W. Souidene, M. Opitz and M. Riemann

Abstract — **This work proposes a system for Acoustic Event Detection and Classification (*AEDC*) using enhanced audio signal provided by a CMT (Coincidence Microphone Technology) microphone. The CMT microphone through signal processing algorithm provides an enhanced signal in several azimuths with a step of 15°. The AEC module exploits this technology to increase classification performance. The automatic detection system based on DWT uses an adaptive threshold for a different energy level and sampling rate quality. The classification system is based on an unsupervised order estimation of Gaussian mixture model adapted to the variability of sound event acoustic information and the representation cost.**

I. Introduction

Audio based surveillance systems stem from the field of automatic audio classification and matching. Traditional tasks in this area are speech/music segmentation and classification or audio retrieval. More recently, specific algorithms covering the detection of particular classes of events for multimedia-based surveillance have been developed.

Acoustic Event Detection and Classification is a recent sub-area of computational auditory scene analysis [1] where particular attention has been paid to automatic surveillance systems [2], [3], [4]. In particular, the use of audio sensors in surveillance and monitoring applications has proven to be particularly useful for the detection of distress situation events, chiefly when the patients suffer from cognitive illness. The recent research work in medicine has concluded that some patients with mild cognitive impairment will develop Alzheimer in the future. The efficient detection and recognition of the distress situation is one part of the socially assistive robotics technology [5] aimed at providing affordable personalized cognitive assistance.

This work deals with the classification of speech and non-speech events, where the considered non-speech events are typical sounds that may occur in everyday life. In practice some of the sound events may be considered as a noise of everyday life which can perturb the recognition task.

The proposed implementation is based on a hierarchical approach that has also been employed in [6]. We propose a specific system able to detect a speech utterance used as input for distress expression recognition system or/and dialogue system. The use of an acoustic system for tracking and recognition remains most useful compared to video surveillance, especially in a home environment. Mainly we consider the human solo sounds as a vital signals like "Snore, Cough, Cry,…,etc.".

We extend the previous work from using an omnidirectional microphone-based, firstly, to exploit the acoustic diversity observed by a set of CMT microphones-based placed far from each other and, secondly, to decrease the mismatch that can be caused by several factors. The aim is to select a useful signal component out of several events occurring at the same time. The CMT microphone localizes the sound event and can provide an enhanced signal if two sound sources are presented at the same time. The main goal is to develop a system that is robust to the presence of noise that might be generated for example by the hairdryer, vacuum cleaner or water flushing.

This research is being conducted under the European Project CompanionAble[1] an internationally active group dedicated to carrying out leading-edge research in computer vision and signal processing for man-machine communication, including patient home-care, gesture-based interaction, biometry, video surveillance.

II. Context and Goals

The proposed audio based surveillance system is developed in the framework of CompanionAble project with the three goals: patient security, domotic application and context awareness.

In order to assure theses goals the global system is designed to use a multiple microphones in each area depending on the room dimensions and properties. The larger room will be equipped with one or two CMT microphones which allow sound localization, however the other rooms will contain omnidirectional microphones. Fig.1 presents the sound processing architecture.

The analysis system consists of the two modules that allow the localization of useful event audio segment. The identification of the event given by the audio segment is carried out on 24 channels generated by a process provided by the CMT microphone. However, the segmentation module is carried out only on the omnidirectional signal. In the case of simultaneous detections the low level data fusion chooses signals based on the signal-to-noise ratio (SNR). The detection module associated with the CMT microphone

J.E. Rougui, D. Istrate and W. Souidene, are with LRIT-ESIGETEL, 1, Rue du Port de Valvins, 77210 Fontainebleau-Avon Cedex, France, {jamal.rougui,dan.istrate,wided.souidene}@esigetel.fr.
M. Opitz and M. Riemann are with AKG Acoustics GmbH Lemböckgasse 21-25 A-1230 Vienna, Austria. martin.opitz@harman.com, marco.riemann@harman.com.

[1] www.companionable.net

communicates with localization algorithm in order to enhance the signal in the useful direction.

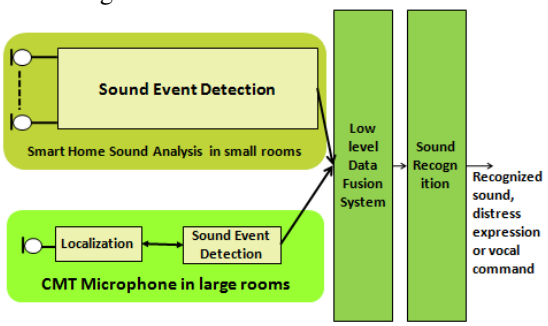

Fig.1 – Sound Processing Architecture

III. CMT AND LOCALIZATION

In order to increase system robustness for all possible locations of the acoustic source, a CMT microphone network is adopted here. At least one microphone per room is used in order to ensure a good spatial coverage

The CMT microphone consists of one pressure transducer and three first order pressure gradient electret transducers, each with a diaphragm, with each pressure gradient transducer having a first sound inlet opening, which leads to the front of the diaphragm, and a second sound inlet opening, which leads to the back of the diaphragm. Both sound inlets are on the same side of the disc shaped pressure gradient transducers.

The three pressure gradient transducers lie all in one plane. Their respective main directions – the directions of their maximum sensitivity – are lying in the same plane and are inclined relative to each other by 120 degrees. The acoustical centers of all 4 microphones are lying close together within a sphere with few millimeters radius.

In the further context we will refer to azimuthal detection of the direction of sound incidence only, as this is the most important localization information in the context of CompanionAble.

IV. SOUND DETECTION AND CLASSIFICATION

The sound flow provided by the CMT microphone is analyzed trough a hierarchical approach that involves firstly a useful signal detection followed by an event classification.

The first sound analysis module is the event detection module which is an important step before the event classification, especially when the events detection occurs in a variable noise of the home environment.

The signal classification starts with a sound/speech identification followed by a classification adapted to the identified signal. If the label was speech, a speech recognition engine is used and if a sound was identified a sound classes recognition system is launched. In this paper we are focusing on the sound identification.

A. *Sound event segmentation*

The audio segmentation must be able to detect a short event like an impulsive signal. Ideally, the segmentation module must be robust against a low signal energy due to a distant acquisition and different acquisition qualities. The classic techniques of event detection are based on the signal energy threshold or on other statistical features threshold [6],[7].

The wavelet based event detection algorithm proposed in [8] was adopted in this work. This algorithm is based on DWT (Discrete Wavelet Transform) using Daubechies wavelets with 6 vanishing moments. An adaptive threshold, depending on average and standard deviation of the energy is applied on the high frequency wavelet transform coefficients.

B. *Unsupervised Gaussian mixture modeling*

The extracted signal is analyzed by a hierarchical classification system. Firstly a classification between vocal and non vocal is carried out. In the case of non vocal signal a new classification between some everyday life sounds and noises is started. The sound classes were defined by CompanionAble consortium in order to allow the distress situation detection but also to help context awareness identification. Each classification module is based on GMM with an optimized number of Gaussian mixture [11]. Fig. 2 presents the hierarchical signal classification and the detected sound classes.

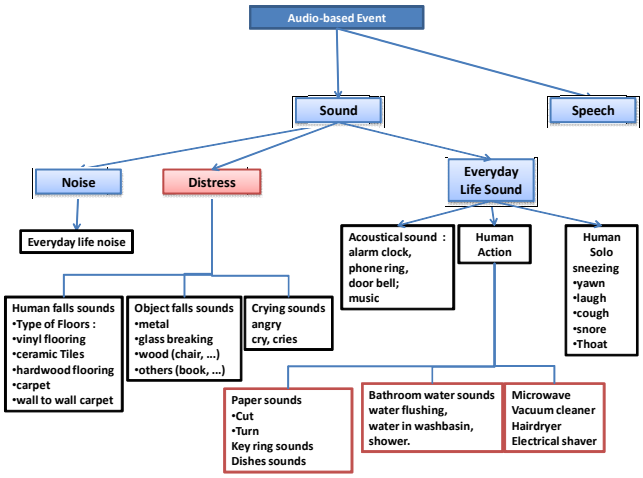

Fig. 2 - Hierarchical Sound event for smart home application

C. *Coupling CMT with sound analysis module*

The sound source localization algorithm of CMT allows listening in 24 directions (15° angular resolution in the horizontal plane). The signal coming from the omnidirectional microphone which contains all information is analyzed by the sound segmentation module. The start and stop information for each detected signal is used in order to segment the 24 azimuthal files. As shown in Fig.3 the processed files given to each azimuth have the same content with a different SNR. The low level data fusion (Fig. 1) is composed by a matching algorithm between all extracted signals in order to choose that one, which is best suited for the classification. In fact the classification is carried out on all segments and the output is a matrix composed by the most probable classification hypothesis for each segment on each azimuth coupled with its likelihood ($ClassHyph_{i,j}$ $ML_{i,j}$). For each detected segment the classification

hypothesis with the Maximum likelihood is considered like the identified signal.

In the next section we compare the results obtained with the omnidirectional microphone, being part of the CMT microphone, with those results obtained on enhanced localized signals.

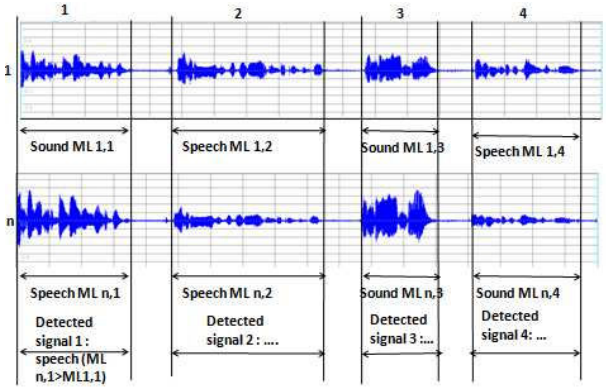

Fig.3 - Real time analysis of audio files processed-based CMT microphone by each azimuth in case of everyday life situations, discussion between 2 persons.

V. EVALUATION

A. Training Corpus for localization and classification

In order to use the localization ability of the CMT microphone, a database with training data has to be recorded in advance. For this purpose, the CMT microphone was placed on an acoustic boundary layer as is also foreseen in the actual application. The impulse responses of all 4 single transducers of the CMT microphone were measured for 24 directions in the horizontal plane corresponding to an azimuth distance of 15 degrees between the single measurements. For the measurements an AKG proprietary PC based measurement system was used. For the measurements, a Tannoy loudspeaker Sytem600 emitting a periodic noise signal with low crest factor was used. Applying the DFT (Discrete Fourier Transform), the corresponding transfer functions were determined and the results were stored in the database of the training corpus. The influences of the measurement loudspeaker, the amplifier and A/D-D/A-converters were determined by a reference measurement with a ½" calibrated measurement microphone and were removed from the CMT microphone data.

The sound classification module has currently 24 sound classes trained on 108' of signal and 5 noises of everyday life (Vacuum cleaner, Water flushing, Dishwasher, HairDrayer, RadioTv) trained on 18' of signal. The classes referring to sound and speech for the first classification level were trained on all existing sounds and on 38' of speech respectively.

B. Test Corpus

In order to evaluate our CMT based sound analysis approach we have recorded 20 scenarios in ESIGETEL laboratory using two CMT microphones. The sound signals were acquired with a RME DSP Multiface II card at 44.1 kHz sampling rate. Calibration of recording level was done using a Tannoy Precision 6D loudspeaker generating white noise with 70dB$_{SPL}$ linear weighting.

The recordings were made in two different rooms: one with a rectangular shape (Fig. 4) and another one with a triangular shape in order to evaluate also the influence of sound reflection on localization and classification.

The 20 scenarios were composed of 10 normal scenarios (everyday life situations, discussion between 2 persons…) and 10 distress scenarios containing the fall of a person (simulated by the actor), some distress expressions or distress sounds. Each scenario has been played in the two rooms and a video recording has been made for easy labeling. The data base has about 34 minutes of recordings.

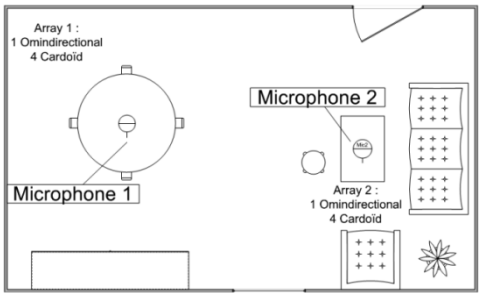

Fig. 4 - CMT microphone layout in rectangular room

C. Results

The proposed system was evaluated on the presented corpus in terms of localization, sound event detection and classification. We present here the example of a Normal Scenario were two persons have a discussion.

1) Localization

For the localization of a sound source with unknown position recorded by the CMT microphone the following strategy is used: the incoming signal is split into blocks of about 20ms length. For each block, a DFT is applied to all 4 signals. The amplitude spectra of the signals stemming from the three pressure gradient transducers are normalized by dividing them by the spectrum of the omnidirectional microphone. Comparing these normalized spectra with all spectra of the database, the direction for the most probable sound incidence is derived. The algorithm used is based on the method described in [10].

In Fig. 5 an example of the localization is shown for a dialogue of 2 speakers. The dialogue shown in Fig. 5 is part of the Normal Scenario 1 listed in Table I. The two speakers were localized such that their voices impinged on the microphone from 300 degrees and 50 degrees azimuth respectively. First a manual tagging of the respective speaker was done. In Fig. 5 the result of the manual tagging is shown with a dashed red line. After the end of each speech section the tagged angle was kept on the last detected azimuth. The automatic detection of the direction of speech sound incidence is shown by the red line in Fig. 5. Apart from minor delays in the attack phase at the beginning of the speech sections the congruence between manually and automatically detected direction of speech sound incidence is very good.

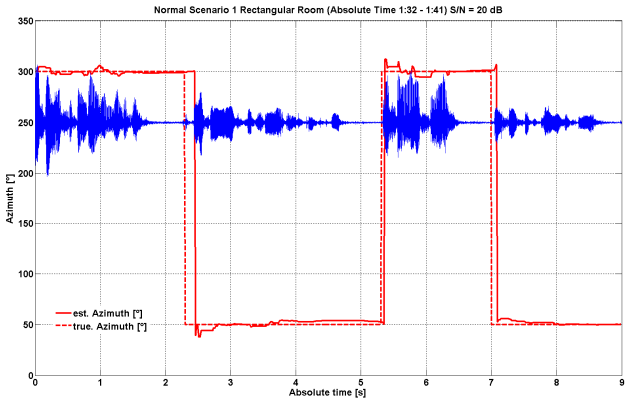

Fig.5 - Localization result for 2 persons speaking

2) Sound detection and classification

The sound event detection is evaluated in terms of number of correct detected events. The recorded signal was manually labeled in SAM format [9]. 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:

$$AED = \frac{N°\ correct\ detected\ events}{N°\ detected\ events} * 100$$

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

$$AEC = \frac{N°\ correct\ classified\ detected\ events}{N°\ correct\ detected\ events} * 100$$

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

$$AEDC = \frac{N°correct\ classified\ events}{N°\ detected\ events} * 100$$

Firstly the proposed sound analysis system is evaluated on the omnidirectional microphone signal, which acquires the signal from all directions. These results are compared with the results obtained from the 24 directions localization files. The analysis is performed on a normal scenario with duration of about 2 minutes (see Table I).

In the Table II we can observe that the classification error rate (1-AEC) decrease from 26.7% in the case of omnidirectional microphone to **11.8%** in the case of the data fusion between different azimuth localization. This can be explained by the fact that SNR is enhanced for some events in some directions (Fig.3).

VI. Conclusion

In this paper we have presented a first approach of an audio based surveillance system for distress situation identification, vocal commands and context awareness detection which was developed in the framework of CompanionAble project. The current proposition uses a CMT microphone which allows localizing the sound source and to enhance the signal. Our first proposition based on the data fusion between different classifications of the same sound event indicates good performances and encourages us to evaluate them on a larger data base.

Acknowledgment

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

Table I
Normal Scenario 1

Time	Duration	Action
00:00	00:20	person is sitting and reading a book
00:20	00:03	person moves the chair & stands up
00:23	00:20	person walks around
00:43	00:03	person sits down again
00:46	00:15	person is reading a book
01:01	00:20	another person is entering the room and is walking around
01:21	00:15	the person is sitting down to the desk
01:36	01:00	the two persons are talking
02:36	00:15	the first person leaves the room

Table II
Detection and classification on Normal Scenario 1

Signal type	AED	AEC	AEDC
Omnidirectional	66.7 %	73.3 %	48.9 %
Fusion on 24 localization signals	66.7 %	88.2 %	60.2 %

References

[1] G. Valenzise, L. Gerosa, M. Tagliasacchi,. F. Antonacci, A. Sarti, Advanced Video and Signal Based Surveillance, 2007. *AVSS 2007.* Volume 2, Issue, 5-7 Sept. 2007 Page(s):21 - 26

[2] D. Wang, G. Brown, "Computational Auditory Scene Analysis: Principles, Algorithms and Application", *Wiley-IEEE* Press, 2006.

[3] C. Zieger and M. Omologo, "Acoustic event classification using a distributed microphone network with a GMM/SVM combined algorithm", *Interspeech*, Brisbane, September 2008, pp. 115-118.

[4] D. Feil-Seifer, and M.J. Matarić, "Defining socially assistive robotics," in Proc. *IEEE International Conference on Rehabilitation Robotics (ICORR'05),* Chicago, Il, USA, June 2005, pp. 465–468.

[5] J.L. Rouas, J. Louradour, and S. Ambellouis. "Audio Events Detection in Public Transport Vehicle". Proc. *of the 9th International, IEEEConference on Intelligent Transportation System 2006, 17-20 Sept. 2006,pp.733 - 738*

[6] T. Yamada, N. Watanabe, F. Asano, N. Kitawaki, "Voice activity detection using non-speech models and HMM composition," Proc. *Workshop on Hands-free Speech Communication*, Apr. 2001, pp. 131–134, Tokyo, Japan, April 2001.

[7] A. Dufaux, "Detection and recognition of Impulsive Sounds Signals," Ph.D. dissertation, Faculté des sciences de l'Université de Neuchâtel, Switzerland, 2001.

[8] D. Istrate, E. Castelli, M. Vacher, L. Besacier, J.-F. Serignat, " Information extraction from sound for medical telemonitoring" *IEEE Transactions on Information Technology in Biomedicine,* Volume 10, Issue 2, April 2006 Page(s):264 – 274.

[9] D. Well, J. Barry, W. Grice, M. Fourcin, and A. Gibbon, SAM ESPRIT PROJECT2589-multilingual speech input/output assessment, methodology and standardization. *University College London*: Final report. *Technical Report SAM-UCLG004*.

[10] K. Freiberger, A. Sontacchi, M. Opitz, *Acoustic source localization using coincident microphone arrays,* Applied for Proc. of the 12th Int. Conference on Digital Audio Effects (DAFx-09), Como, Italy, September 11-4, 2009.

[11] Bouman97, "Cluster: An unsupervised algorithm for modeling Gaussian mixtures", available from *http://www.ece.purdue.edu/~bouman* , April, 1997.