Fear-type emotion recognition and abnormal events detection for an audio-based surveillance system

C. Clavel & T. Ehrette
Thales Research and Technology France, RD 128,
91767 Palaiseau Cedex, France

Abstract

The goal of our research work is to carry out an audio-based abnormal situation diagnosis system for the SERKET project which aims at developing surveillance systems dealing with dispersed data coming from heterogeneous sensors. We look at things from the point of view of human life protection in the context of civil safety. Therefore, we focus on abnormal situations during which human life is threatened (psychological and physical attack). The proposed system relies on information conveyed by both speech and non-speech acoustic manifestations to generate alerts. More precisely, our audio module can be divided into two sub-modules. The first one concerns the abnormal event detection system that is illustrated here in the case of gun-shot. The second one focuses on information conveyed by the emotional content of speech. Such information is useful for decoding human behaviour in abnormal situations and so provides a situation diagnosis. More precisely the targeted emotions are fear-type emotions corresponding to symptomatic emotions occurring when the matter of survival is raised, including the different fear-related emotional states from worry to panic. At the last stage of its development, this work would propose a surveillance system plugged into a real control room. Thus, we proposed here a mock-up to illustrate the running of these two systems.

1 Introduction

Civil safety has received growing interest over the last few years. Recent projects such as SERKET (http://www.research.thalesgroup.com/software/cognitive_solutions/Serket/index.html) tackle the issue of the development of automatic

WIT Transactions on Information and Communication, Vol 39, © 2008 WIT Press
www.witpress.com, ISSN 1743-3517 (on-line)
doi:10.2495/RISK080461

surveillance systems. Such systems aim at facilitating the monitoring of alerts to human workers. The parallel surveillance of multiple screens indeed increases the cognitive overload of the staff and raises the matter of vigilance.

Existing automatic surveillance systems are essentially based on video cues to detect abnormal situations: intrusion, abnormal crowd movement, etc. The trend is now to integrate information obtained from other sensors – audio sensors in particular – into surveillance systems [1–3]. Using several sensors increases the available information and strengthens the quality of the abnormal situation diagnosis. In particular, audio information is useful when the abnormal situation manifestations are poorly expressed by visual cues such as gun-shot events or human shouts or when these manifestations go out of the cameras' shot or occur in the dark.

At the same time, audio event classification and detection have had an increase in interest especially in the context of audio retrieval and indexing [4–6]. Techniques used in this context demonstrate the scientific progress made in recnet years in audio processing. This scientific progress is accompanied by the sophistication of the technology available in audio capture [3].

The goal of this paper is to carry out an audio-based abnormal situation diagnosis system. We look at things from the point of view of human life protection in the context of civil safety. Therefore, we focus on abnormal situations during which human life is threatened (psychological and physical attack). The proposed system relies on information conveyed by both speech and non-speech acoustic manifestations to generate alerts. More precisely, our audio module can be divided in two sub-modules:

The first one concerns the abnormal event detection system which is illustrated here in the case of gun-shot.

The second one focuses on information conveyed by the emotional content of speech. Such information is useful to decode human behaviors in abnormal situations and so to provide a situation diagnosis.

The two next sections present each sub-module of the audio processing module of the dedicated surveillance platform. The running of these two systems is illustrated in a mock-up presented in the last section before conclusions and perspectives.

2 Abnormal audio event detection

In the surveillance context, one of the major difficulties of an audio detection system is linked to the environmental noise that is often non stationary and that may be loud compared to the audio event to detect. We propose here an approach to develop the detection of an event in an audio stream, in noisy conditions. Although our event detection system is currently limited to gun-shot detection, the methodology and the approach followed in this system could be extended to other classes of characteristic sounds of abnormal situations in a given environment.

The gun-shot detection system presented in this paper is based on a novelty detection approach [7]. Indeed novelty detection is based on the evaluation of a

distance to a *normal* situation. This normal situation is built on acoustic data of a given environment.

2.1 Synopsis

The input audio stream is segmented into successive segments which are labeled according to the two main classes (the *shot* class and the *normal* class that represents the typical environment acoustic characteristics). As depicted on figure 1, the architecture of our audio event detection system includes a feature extraction module, a training module that is used to build the model of the two classes using Gaussian Mixture Models (GMM) and a classification module that labels the successive audio segments. The input audio stream is first segmented into short frames (20 ms) and a label is given for segments of 0.5 second (with 50% overlap). Acoustic features extraction, training and detection steps are detailed in [1].

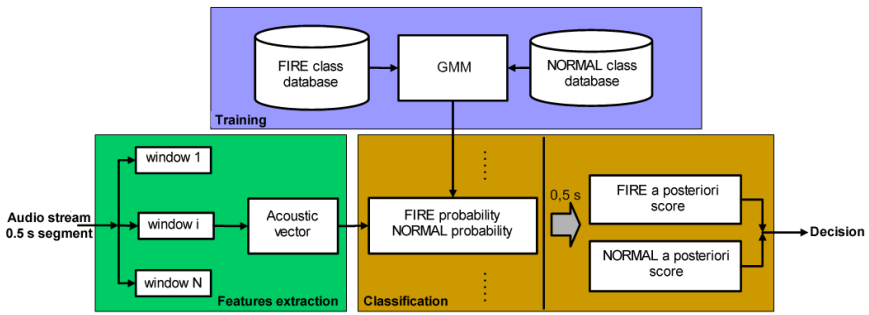

Figure 1: Gun-shot detection system.

2.2 Database and protocols

Corpora of typical audio events in ecological conditions, such as surveillance applications, are not available mainly because of the confidential nature of the data but also because abnormal situations happen rarely and are thus difficult to record. To be as close as possible to real conditions, we gathered artificial data from a set of market places and gun shots recordings extracted from a CD of sounds for the national French public radio [8].

For every four market places recording (*surrounding sequence*) the last 75 seconds are kept for the *normal* class training. The rest of the environmental database is used to build the test database. The test database results from a mix between the gun-shots and the *surrounding sequences*. A shot occurs in each sequence at a random moment with various local Signal-to-Noise Ratio (SNR). The SNR is computed for the part of the surrounding sequence where the shot is inserted and data are previously normalized before mixing. Each test sequence is

30 seconds long and is randomly chosen among test part of market surrounding sequences. For each SNR (from 20 to 5 dB) 134 sequences totaling about 67 minutes are generated for the test. Such mixed test sequences provide a simulation of abnormal situation in a public place as close as possible to the reality (in the case of gun shot). Despite their artificial nature these sequences allow us to control the SNR and therefore to assess the system noise robustness but also to have a ground truth annotation of the test files (i.e. an exact localization of all shot events in the *surrounding sequences*).

We use a *leave one shot out* cross validation method for the training of the *shot* class: every shot to be detected in the test database is removed from the training database during the training step for each test sequence.

For different SNR conditions, the labels given by our automatic event detection system are compared to the ground truth annotation. The overall results are given by computing a false rejection ratio and a false detection ratio.

2.3 Experiments and results

A first evaluation of the system performance is carried out using the training shot database described in the section 2.2. In order to improve tbe system performance, we chose to train acoustic models of shots in an environmental noise. For the *shot* class, a database of shots mixed with surrounding sequences segments is generated from 134 initial shots. Shots are inserted for different SNRs going from 20 to 5 dB with 5dB step. Gun-shot detection is assessed for each SNR level of the test database and of the training database. We present here the major results already detailed in [1].

As expected results rapidly degrade when the SNR condition of the test sequences decreases, i.e. when the shot energy decreases compared to the *surrounding sequences* energy. In particular clean shot (training) databases provide insufficient results in terms of false rejection for the noisiest test sequences. However, we observed that the use of too noisy shot training databases triggers off a considerable increase of false detection rate which reaches 43% in the worst case (5 dB SNR training database and 5 dB SNR test sequences condition). Best results are obtained in all test conditions with a 20 dB SNR training database: false rejection less than 11% and false acceptation less than 15%.

3 Fear-type emotions detection

Recent research on emotional speech reveals the need to consider the emotional content of speech because it influences the semantic decoding of human behavior. For example, in dialog systems applications the determination of the speaker's emotional state aims at adapting the dialog strategy in order to provide a more relevant answer to the speaker's request [9].

We address here the new issue of exploiting the speech emotional component in surveillance systems. The targeted emotions are fear-type emotions corresponding

to symptomatic emotions occurring when the matter of survival is raised, including the various fear-related emotional states [10] from worry to panic.

3.1 Synopsis

The fear-type emotions detection system focuses on differentiating *Fear* class from *Neutral* class. The *Fear* class gathers all fear-related emotional states and the *Neutral* class corresponds to non-negative and non-positive emotional speech with a faint emotional activation. The audio stream has been manually pre-segmented into decision frames, called *segments* which correspond to a speaker turn or a section of speaker turn portraying the same annotated emotion. The system is based on acoustic cues and focuses as a first step on a classification of the predefined emotional segments.

The classification system merges two classifiers, the *voiced classifier* and the *unvoiced classifier* which consider respectively the voiced portions and the unvoiced portions of the segment [11]. The emotional manifestations conveyed by unvoiced speech portions need also to be modeled. Emotions in abnormal situations are indeed accompanied by a strong body activity, such as running or tensing, which modifies the speech signal, by increasing the proportion of unvoiced speech in particular.

The first step of the overall system aims at extracting prosodic, voice quality and spectral features [11]. The second step aims at reducing the feature space using the Fisher selection algorithm in two steps in order to avoid strong redundancies. The third step consists in the training of the models for each voicing condition (using GMM). The final step consists in the classification of each segment according to the two main classes (the *fear* class and the *neutral* class) by merging the results of the two classifiers (voiced or unvoiced). The weight is assigned to the classifiers depending on the proportion of voiced frames in the segment. For more details about the overall system see [12].

3.2 Evaluation material

Existing real-life corpora [10] illustrate everyday life contexts in which social emotions currently occur. The type of emotional manifestations and the degree of intensity of such emotions are determined by politeness habits and cultural behaviours. The emotions targeted by surveillance applications belong to the specific class of emotions emerging in abnormal situations. They occur indeed in dynamic situations, during which the matter of survival is raised. Abnormal situations are however rare and unpredictable and real-life recordings of such situations are often confidential.

We built the SAFE Corpus (Situation Analysis in a Fictional and Emotional Corpus) in order to provide an estimation of emotion acoustic particularities of fear-type emotions in abnormal situation [13]. The SAFE Corpus consists of audio-visual sequences (a total of 7 hours) extracted from a collection of 30 recent movies in English language. Such fiction movies provide an interesting range

of potential real-life abnormal contexts and of type of speakers that would have been very difficult to collect in real life. Emotions are considered in their temporal context. We segmented each sequence that provides a particular context into a basic annotation unit, the *segment*, which has been defined in Section 3.1. A total of 5275 segments of speech with a duration varying from 40ms to 80s are thus obtained from the 400 sequences of the corpus.

A rich annotation strategy was developed and takes into account various aspects of the sequences content [13]. The *emotional substance* is considered at the segment level and includes among other descriptors a description in four major emotion classes: *Fear*, *Other Negative Emotions*, *Neutral*, *Positive Emotions*. The *context of emergence* is also described by a threat track, a speaker track and an annotation of audio environment. Two labelers annotated the corpus. The kappa score [14] between the two labelers is at 0.49 which is an acceptable level of agreement for subjective phenomena such as emotions.

The following experiment and analysis are performed on a subcorpus containing only *good quality* segments labeled *Fear* and *Neutral*. The quality of the speech in the segments concerns the speech audibility and has been evaluated by the labelers. Remaining segments include various environment types (noise, music). Overlaps have been avoided. Only segments where the two human labelers agree are considered, i.e. a total of 994 *segments* (38% of *Fear* segments and 62% of *Neutral* segment).

The test protocol follows the *Leave One Movie Out* method: the data is divided into 30 subsets, each subset contains all the segments of a movie. 30 trainings are performed, each time leaving out one of the subsets from training and using only the omitted subset for the test. This protocol ensures that the speaker used for the test is not found in the training database.

3.3 Results

Best results are obtained when the unvoiced classifier is considered with a weight decreasing quickly when the voiced rate increases [11]. The Mean Accuracy Rate (MAR) is 70.8%. The confusion matrix resulting from the *Fear* vs. *Neutral* classifier is presented in Table 1. It illustrates the confusions between the automatic labeling of the classifier and the manual labels provided by the labelers. With regard to the fear recognition, 70.3% of the segments labelled *Fear* are correctly recognized by the system. The local system behavior on the various segments according to the threat during which they occur is detailed in [12].

4 Abnormal situation detection mock-up

At the last stage of its development, this work would propose a surveillance system plugged in a real control room. To have a more precise view on how the final system would work, we proposed a mock-up. It combines two detectors, one for gun-shot and one for fear-type emotion detection. They communicate together on a multi-agent platform named OAA that stands for Open Agent Architecture

Table 1: Confusion matrix in percent of the *Fear* vs. *Neutral* classification system.

manual \ automatic	Neutral	Fear
Neutral	**71.3**	28.7
Fear	29.7	**70.3**
Mean Accuracy Rate	**70.8**	

[15]. This solution allows to distribute the computation workload on different servers that share autonomous services. Each service is carried out by a so-called textitagent. All the agents communicate together via a facilitator centralizing the exchanges. Four agents compose our mock-up:

1. sequence acquisition
2. gun-shot detection
3. fear detection
4. results presentation

The last agent is the interface: a PC screen split into three frames (see 2). The top frame is for the video. A movie sequence can be chosen, loaded and played on this area. When the sequence starts, the audio stream, displayed in the center frame, is computed in parallel by the two detection agents. A response occurs at the end of each segment. A rectangular box appears in the bottom frame and contains the time codes of the segment and the answer ("gun-shot", "fear" or "neutral"). At the end of the whole movie sequence, the bottom frame is composed by many boxes of different colors. This simplify the assessment of the system: red boxes correspond to missed detection or false alarm, blue boxes to correct answers (according to the human annotations) and gray boxes to unprocessed segments because of overlap or very poor recording quality.

As defined in section 3.1 a segment is a speaker turn portraying one single emotion. Our preliminary system is based on a manually segmentation, we have planned to replace it by an automatic segmentation based on techniques such as vocal activity detection, speaker tracking, etc.

5 Conclusion and perspectives

In this paper we proposed a novel idea of public places surveillance by focusing on audio information. This idea has been implemented and proved its efficiency and added value for this kind of application. Our approach combines an audio event detection system with an emotion detection system. We selected acoustic features that are highly correlated with fear manifestation, such as the voiced trajectory duration, the voiced/unvoiced coefficient or spectral sub-band energy. Results are encouraging: more than 70% of the corpus sequences of fear are correctly identified as so. But two facts must be taken into account to balance this result:

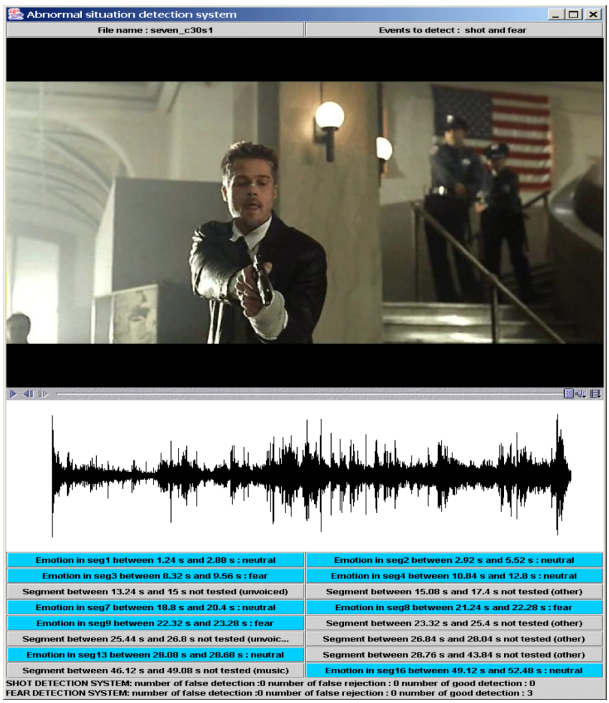

Figure 2: Mock-up layout example.

1. It is not performed on real data but on action movie sequences: the variability of the recording quality (kind of microphone, position), of environmental acoustic features (reverberation) or of kind of actions is higher than in real-life.
2. Segments are annotated by human labelers. They have a personal feeling that can modify their judgment of the situation. Thus the system can be confused in the learning phase and in the evaluation phase because its results have no strong reference that could be given by an expert.

In future works we would like to extend the number of targeted emotions and refine the concept of fear by distinguishing the context of emergence: imminent, present and past, and to be able to qualify the emotion in intensity. To enhance the results one thing should also be taken into account: the acoustic environment of the place were data are collected. An underground car park presents a much higher reverberating coefficient than a market place and algorithms can be confused if different situations are mixed.

Finally, more than a surveillance application, we proposed an overall methodology to go from data (format, annotation) to a final system by building the appropriate models to get the application ready to be used in the real world.

References

[1] Clavel, C., Ehrette, T. & Richard, G., Events detection for an audio-based surveillance system. *International Conference on Multimedia and Expo (ICME)*, Amsterdam, Netherlands, 2005.

[2] Atrey, P., Maddage, N. & Kankanhalli, M., Audio based event detection for multimedia surveillance. *Proc. of ICASSP*, 2006.

[3] Smeaton, A.F. & McHugh, M., Towards event detection in an audio-based sensor network. *VSSN '05: Proceedings of the third ACM international workshop on Video surveillance & sensor networks*, ACM Press: New York, NY, USA, pp. 87–94, 2005.

[4] Cai, R., Lu, L., Zhang, H.J. & Cai, L.H., Highlight sound effects detection in audio stream, 2003.

[5] Pfeiffer, S., Fischer, S. & Effelsberg, W., Automatic audio content analysis. *Proc. of ACM international conference on Multimedia*, pp. 21–30, 1997.

[6] Essid, S., Richard, G. & David, B., Instrument recognition in polyphonic music based on automatic taxonomies. *IEEE Transactions on Speech and Audio Processing*, 2006.

[7] Markou, M. & Singh, S., Novelty detection: a review. *Signal Processing*, **83(12)**, pp. 2481–2497, 2003.

[8] Mercier, D., Sound library. Audivis Distribution, 1989. CD.

[9] Devillers, L. & Vasilescu, I., Prosodic cues for emotion characterization in real-life spoken dialogs. *Eurospeech*, Genve, 2003.

[10] Cowie, R. & Cornelius, R., Describing the emotional states that are expressed in speech. *Speech Communication*, **40**, pp. 5–32, 2003.

[11] Clavel, C., Vasilescu, I., Richard, G. & Devillers, L., Voiced and unvoiced content of fear-type emotions in the safe corpus. *Speech Prosody*, Dresden, Germany, 2006.

[12] Clavel, C., Devillers, L., Richard, G., Vasilescu, I. & Ehrette, T., Abnormal situations detection and analysis through fear-type acoustic manifestations. *Proc. of ICASSP*, Honolulu, 2007. In press.

[13] Clavel, C., Vasilescu, I., Devillers, L. & Ehrette, T., Fiction database for emotion detection in abnormal situations. *International Conference on Spoken Language Processing (ICSLP)*, Jeju, Korea, 2004.

[14] Craggs, R., Annotating emotion in dialogue - issues and approaches. *Proc. of CLUK Research Colloquium*, 2004.

[15] Cheyer, A. & Martin, D., The open agent architecture. *Journal of Autonomous Agents and Multi-Agent-System*, **4(1)**, pp. 143–148, 2001.