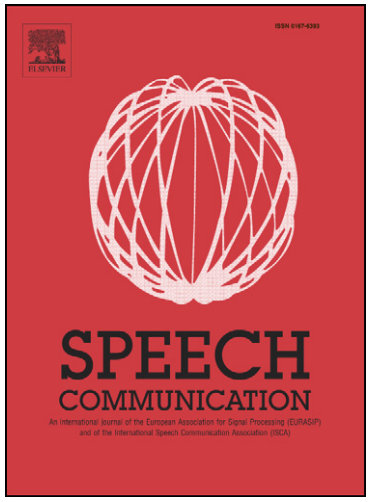
Fear-type emotion recognition for future audio-based surveillance systems

Chloé Clavel, I. Vasilescu, L. Devillers, Gael Richard, T. Ehrette



Fear-type emotion recognition for future audio-based surveillance systems

C. Clavel, I. Vasilescu, L. Devillers, G. Richard, T. Ehrette

PII: S0167-6393(08)00037-X
DOI: 10.1016/j.specom.2008.03.012
Reference: SPECOM 1696

To appear in: *Speech Communication*

Received Date: 26 October 2007
Revised Date: 14 March 2008
Accepted Date: 14 March 2008

Please cite this article as: Clavel, C., Vasilescu, I., Devillers, L., Richard, G., Ehrette, T., Fear-type emotion recognition for future audio-based surveillance systems, *Speech Communication* (2008), doi: 10.1016/j.specom.2008.03.012

Fear-type emotion recognition for future audio-based surveillance systems

C. Clavel [a] I. Vasilescu [b] L. Devillers [b] G. Richard [c] T. Ehrette [a]

[a] *Thales Research and Technology France, RD 128, 91767 Palaiseau Cedex, France*
[b] *LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France*
[c] *TELECOM ParisTech, 37 rue Dareau, 75014 Paris, France*

Abstract

This paper addresses the issue of automatic emotion recognition in speech. We focus on a type of emotional manifestation which has been rarely studied in speech processing: fear-type emotions occurring during abnormal situations (here, unplanned events where human life is threatened). This study is dedicated to a new application in emotion recognition - public safety. The starting point of this work is the definition and the collection of data illustrating extreme emotional manifestations in threatening situations. For this purpose we develop the SAFE corpus (Situation Analysis in a Fictional and Emotional corpus) based on fiction movies. It consists of 7 hours of recordings organized into 400 audiovisual sequences. The corpus contains recordings of both normal and abnormal situations and provides a large scope of contexts and therefore a large scope of emotional manifestations. In this way, not only it addresses the issue of the lack of corpora illustrating strong emotions, but also it forms an interesting support to study a high variety of emotional manifestations. We define a task-dependent annotation strategy which has the particularity to describe simultaneously the emotion and the situation evolution in context. The emotion recognition system is based on these data and must handle a large scope of unknown speakers and situations in noisy sound environments. It consists of a fear vs. neutral classification. The novelty of our approach relies on dissociated acoustic models of the voiced and unvoiced contents of speech. The two are then merged at the decision step of the classification system. The results are quite promising given the complexity and the diversity of the data: the error rate is about 30%.

Key words: Fear-type emotions recognition; Fiction corpus; Annotation scheme; Acoustic features of emotions; Machine learning; Threatening situations; Civil safety.

Email address: chloe.clavel@thalesgroup.com (C. Clavel).

1 Introduction

One of the challenges of speech processing is to give computers the ability to understand human behaviour. The computer input is the signal captured by a microphone, i.e. the low level information provided by audio samples. Closing the gap between this low level data and understanding of human behaviour, it's a scientific challenge. Consequently, the issue now is not only to know what is said but also to know the speaker's attitude, emotion or personality.

This paper concerns the emerging research field of emotion recognition in speech. We propose to investigate the integration of emotion recognition in a new application, namely automatic surveillance systems. This study comes within the scope of the SERKET [1] project, which aims to develop surveillance systems dealing with dispersed data coming from heterogeneous sensors, including audio sensors. It is motivated by the crucial role played by the emotional component of speech in the understanding of human behaviour, and therefore in the diagnosis of abnormal situation.

Our audio-based surveillance system is ultimately designed to consider the information conveyed by abnormal non-vocal events such as gunshots (Clavel et al., 2005), though we focus here on the part of the system dealing with vocal manifestations in abnormal situations. We look at things from the viewpoint of protecting human life in the context of civil safety and we choose to focus on abnormal situations during which human life is in danger (e.g. fire, psychological and physical attack). In this context, the targeted emotions correspond to a type of emotional manifestation which has been so far rarely studied – fear-type emotions occurring during abnormal situations.

The development of an emotion recognition system can be broken down into four distinct steps: the acquisition of emotional data, the manual annotation of the emotional content, the acoustic description of the emotional content and the development of machine learning algorithm. Given that the emotional phenomenon is especially complex and hard to define, these steps require the know-how of a set of distinct disciplines such as psychology, social sciences, biology, phonetic, linguistic, artificial intelligence, statistics, acoustics and audio signal processing. In this introduction, we set out first to unravel the know-how of these disciplines from an emotion recognition system point of view and then to present the additional challenges implied by the surveillance application.

[1] http://www.research.thalesgroup.com/software/cognitive_solutions/Serket/index.html

1.1 Overview of emotion recognition systems

1.1.1 Acquisition of the emotional recordings

The basis of emotion research studies is the acquisition of data that are recordings of emotional manifestations. More precisely, data are required for the conception of emotion recognition systems so that the machine can learn to differentiate the acoustic models of emotion. In this case, the challenge is to collect a large number of recordings illustrating emotions as they are expected to occur in application data. In particular, data should be ideally representative of everyday life if the application has to run in everyday life contexts (Douglas-Cowie et al., 2003). Besides, not only the type of collected emotions but also the type of pictured contexts should be appropriate for the targeted application. The context of emotion emergence concerns the situation (place, triggering events), the interaction (human-human or human-machine, duration of the interaction), the social context (agent-customer for call centres), the speaker (gender, age), the cultural context, the linguistic context (language, dialect), and the inter-modal context (gesture and speech for surveillance applications or speech alone for call centres).

The HUMAINE network of excellence has carried out an evaluation of the existing emotional databases[2]. This evaluation shows that one requirement is not adequately addressed in existing databases: there is a lack of corpora illustrating strong emotions with an acceptable level of realism. Indeed specific real-life emotional data are difficult to collect given their unpredictable and confidential nature. That's the reason why acted databases are still used to a large extent in emotional speech studies: Juslin and Laukka (2003) list 104 studies on emotions and estimate at 87% the percentage of studies carried out on acted data. The difficulty is greater when dealing with extreme emotions occurring in real-life threat contexts and extreme emotions are almost exclusively illustrated in acted databases (Mozziconacci (1998), Kienast and Sendlmeier (2000), van Bezooijen (1984), Abelin and Allwood (2000), Yacoub et al. (2003), McGilloway (1997), Dellaert et al. (1996), Banse and Scherer (1996), Bänziger et al. (2006)).

Acted databases generally tend to reflect stereotypes that are more or less far from emotions likely to occur in real-life contexts. This realism depends on the speaker (professional actor or not) and on the context or the scenario provided to the speaker for emotion simulation. Most acted databases are laboratory data produced under conditions designed to remove contextual information. Some recent studies have aimed to collect more realist emotional portrayals by using acting techniques that are thought to stir genuine emotions through

[2] http://emotion-research.net/wiki/Databases

action (Bänziger et al., 2006), (Enos and Hirshberg, 2006).

An alternative way to obtain realistic emotional manifestations is to induce emotions without speaker's knowledge, such as with the eWIZ database (Aubergé et al., 2004), and the SAL database (Douglas-Cowie et al., 2003). However, the induction of fear-type emotions may be medically dangerous and unethical, so that fear-type emotions are not illustrated in elicited databases.

The third type of emotional database, real-life database, illustrates, to a large extent, everyday life contexts in which social emotions currently occur. Some real-life databases illustrate strong emotional manifestations (Vidrascu and Devillers (2005), France et al. (2003)) but the types of situational contexts are very specific (emergency call centre and therapy sessions), which raises the matter of using databases illustrating a restricted scope of contexts (as defined previously) for various applications.

1.1.2 Annotation of the emotional content

The second step consists of the emotional content annotation of the recordings. The challenge is to define an annotation strategy which is a good trade-off between genericity (data-independent) and the complexity of the annotation task. Annotated data are required not only to evaluate the performance of the system, but also to build the training database by linking recordings to their emotional classes. The annotated data must therefore provide an acceptable level of agreement. However, the emotional phenomenon is especially complex and subjected to discord. According to Scherer et al. (1980), this complexity is increased by the two opposite effects push/pull implied in emotional speech: physiological excitations "push" the voice in one direction and conscious attempts driven by cultural rules "pull" them in an another direction.

The literature on emotion representation models has its roots in psychological studies, and offers two major description models. The first one consists of representing the range of emotional manifestation in abstract dimensions. Various dimensions have been proposed and vary according to the underlying psychological theory. The activation/evaluation space is recently the one which is used the most frequently and is known to capture a large range of emotional variation (Whissel, 1989).

The second one consists of using categories for the emotion description. A large amount of studies dedicated to emotional speech use a short list of 'basic' (see the overview of Orthony and Turner (1990)) or 'primary' (Damasio, 1994) emotion terms which differ according to the underlying psychological theories. The 'Big Six' (fear, anger, joy, disgust, sadness and surprise) defined by Ekman and Friesen (1975) are the most popular. However fuller lists (Ekman (1999), Whissel (1989), Plutchik (1984)) have been established to describe 'emotion-

related states' (Cowie and Cornelius, 2003) and Devillers et al. (2005b) have shown that emotions in real-life are rarely 'basic emotions' but complex and blended emotional manifestations. At a cognitive level, this type of description involves drawing frontiers in the perceptive space. Each emotional category may be considered as a prototype – center of a class of more or less similar emotional manifestations (Kleiber, 1990) which can be linked to other similar manifestations. The difficulty of the categorization task strongly depends on the emotional material. The majority of acted databases aim to illustrate predefined emotional prototypes. All the emotional manifestations illustrated by this type of corpus are strongly convergent to the same prototype. By contrast, emotions occurring in real-life corpora are uncontrolled. They display unpredictable distances to their theoretical prototype. This propensity to occur in different situations through various manifestations engenders labelling challenges when one makes use of a predefined list of labels. In addition, the complexity of emotional categorization is increased by the diversity of the data.

Existing annotation schemes fall short of industrial expectations. Noting this, which is closely akin to the motivation of the EARL proposal [3] by the W3C Emotion Incubator Group, leads us to unravel the emotion description task from an emotion recognition system point view.

1.1.3 Acoustic description of the emotional content

After the emotional material has been collected and labelled, the next step is to extract from the speech recordings acoustic features characterizing the various emotional manifestations. This representation of speech signal will be used as the input of the emotion classification system. Existing representations are based on both high-level and low-level features. High-level features, such as pitch or intensity, aim at characterizing the speech variation accompanying physiological or bodily emotional modifications (Picard (1997) Scherer et al. (2001)). First studies focus on prosodic features which include typically pitch, intensity and speech rate and are largely used in emotion classification systems (Kwon et al. (2003) McGilloway (1997) Schuller et al. (2004)) and stand out to be especially salient for fear characterization (Scherer (2003), Devillers and Vasilescu (2003), Batliner et al. (2003)). Voice quality features which characterize creaky, breathy or tensed voices have also recently been used for emotional content acoustic representation (Campbell and Mokhtari, 2003). Low-level features such as spectral and cepstral features, were initially

[3] Emotion Annotation and Representation Language http://emotion-research.net/earl/proposal. The W3C Emotion Incubator Group, after one year of joint work involving several HUMAINE partners, has published its Final Report and a paper in ACII 2007 (Schroeder et al., 2007)

used for speech processing systems, but can also be used for emotion classification systems (Shafran et al. (2003) Kwon et al. (2003)).

1.1.4 Classification algorithms

The final step consists in the development of classification algorithms which aim to recognize one emotional class among others or to classify emotional classes among themselves. The used emotional classes vary according to the targeted application or the type of studied emotional data.

Emotion classification systems are essentially based on supervised machine learning algorithms: Support Vector Machines (Devillers and Vidrascu, 2007), Gaussian Mixture Models (Schuller et al., 2004), Hidden Markov Models (Wagner, 2007), k nearest neighbors (Lee et al., 2002), etc.. It is rather difficult to compare the efficiency of the various existing approaches, since no evaluation campaign has been carried out so far. Performances are besides not only dependent on the adopted machine learning algorithm but also on:

- the diversity of the tested data: contexts (speakers, situations, types of interaction), recording conditions;
- the emotional classes (number and type);
- the training and test conditions (speaker-dependent or not) (Schuller et al., 2003);
- the techniques for acoustic feature extraction which are more or less dependent of prior knowledge of the linguistic content and of the speaker identity (normalization by speaker or by phone, analysis units based on linguistic content).

A first effort to connect existing systems has been carried out with the CEICES (Combining Efforts for Improving automatic Classification of Emotional user States) launched in 2005 by the FAU Erlangen through the HUMAINE [4] network of excellence (Batliner et al., 2006).

1.2 Contributions

It emerges from the previous overview that the development of emotion recognition systems is a recent research field and the integration of such systems in effective applications requires to raise new scientific issues. The first system on laboratory emotional data was indeed carried out recently by Dellaert et al. (1996). Although some emotion recognition systems are now dealing with spontaneous and more complex data (Devillers et al., 2005b), this research field

[4] http://www5.informatik.uni-erlangen.de/Forschung/Projekte/HUMAINE/?language=en

just begins to be studied with the perspective of industrial applications such as call centres (Lee et al., 1997) and human-robot interaction (Oudeyer, 2003). In this context, our approach contributes to an important challenge, since the surveillance application implies the consideration of a new type of emotion and context – fear-type emotions occurring during abnormal situations – and the integration of new constraints.

1.2.1 The application: audio-surveillance

Existing automatic surveillance systems are essentially based on video cues to detect abnormal situations: intrusion, abnormal crowd movement, etc.. Such systems aim to provide an assistance to human operators. The parallel surveillance of multiple screens increases indeed the cognitive overload of the staff and raises the matter of vigilance.

However audio event detection has only begun to be used in some specific surveillance applications such as medical surveillance (Vacher et al., 2004). Audio cues, such as gun shots or screams (Clavel et al., 2005) typically, may convey useful informations about critical situations. Using several sensors increases the available information and strengthens the quality of the abnormal situation diagnoses. Besides 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 shot of the cameras.

1.2.2 The processing of a specific emotional category: fear-type emotions occurring during abnormal situations.

Studies dedicated to the recognition of emotion in speech commonly refer to a restricted number of emotions such as the 'Big Six' (see 1.1.2) especially when they are based on acted databases. Among the studied emotions, fear-type emotions in their extreme manifestations are not frequently studied in the research field of real-life affective computing. Studies prefer to take into account more moderate emotional manifestations which occur in everyday life and which are shaped by politeness habits and cultural behaviours. Indeed, a large part of applications is dedicated to improve the naturalness of the human-machine interaction for everyday tasks (dialog systems for banks and commercial services (Devillers and Vasilescu, 2003), artificial agents (Pelachaud, 2005), robots (Breazeal and Aryananda, 2002)). However, some applications, such as dialog systems for military applications (Varadarajan et al., 2006), (Fernandez and Picard, 2003) or emergency call centres (Vidrascu and Devillers, 2005), deal with strong fear-type emotions in specific contexts (see Section 1.1.1)

The emotions targeted by surveillance applications belong to the specific class

of emotions emerging in abnormal situations. More precisely fear-type emotions may be symptomatic for threat situations where the matter of survival is raised. Here, we are looking for fear-type emotions occurring in dynamic situations, during which the matter of survival is raised. In such situations some expected emotional manifestations correspond to primary manifestations of fear (Darwin, 1872): they may occur as a reaction to a threat. But the targeted emotional class includes also more complex fear-related emotional states (Cowie and Cornelius, 2003) ranging from worry to panic.

Fear manifestations are indeed varying according to the imminence of the threat (potential, latent, immediate or past). For our surveillance application, we are interested in the human assistance by detecting not only the threat but also the threat emergence. There is therefore a strong interest to consider all the various emotional manifestations inside the fear class.

1.2.3 The application constraints

From a surveillance application point of view, the emotion recognition system has to:

- run on data with a high diversity in terms of number and type of speakers,
- cope with more or less noisy environments (e.g. bank, stadium, airport, subway, station),
- be speaker independent and cope with a high number of unknown speakers,
- be text-independent, i.e. not rely on a speech recognition tool, as a consequence of the need to deal with various qualities of the recorded signal in a surveillance application.

1.2.4 Approach and outline

In this paper, we tackle all the various steps involved in the development of an emotion recognition system:

- the development of a new emotional database in response to the application constraints: the challenge is to collect data which illustrate a large scope of threat contexts, emotional manifestations, speakers, and environments (Section 2),
- the definition and development of a task-dependent annotation strategy which integrates this diversity and the evolution of emotional manifestations according to the situation (Section 2),
- the extraction of relevant acoustic features for fear-type emotions characterization: the difficulty relies in finding speaker-independent and text-independent relevant features (Section 3),

- the development of an emotion recognition system based on machine-learning techniques: the system needs to be robust to the variability of the expected data and to the noise environment (Section 3),
- the performance evaluation in experimental conditions as close as possible to those of the effective targeted application (Section 4).

2 Collection and annotation of fear-type emotions in dynamic situations

2.1 Collection of audiovisual recordings illustrating abnormal situations

Abnormal situations are especially rare and unpredictable and real-life surveillance data are often inaccessible in order to protect personal privacy. Given these difficulties, we chose to rely on a type of support hitherto unexploited by emotional studies, namely the fiction. Our fiction corpus (the SAFE Corpus – Situation Analysis in a Fictional and Emotional corpus) consists of 400 audiovisual sequences in English extracted from a collection of 30 recent movies from various genres: thrillers, psychological drama, horror movies, movies which aim at reconstituting dramatic news items or historical events or natural disasters. The duration of the corpus totals 7 hours of recordings organized in sequences from 8 seconds to 5 minutes long. A sequence is a movie section illustrating one type of situation – kidnapping, physical aggression, flood etc. The sequence duration depends on the way of illustration and segmentation of the targeted situation in the movie. A majority – 71 % – of the SAFE corpus depicts abnormal situations with fear-type emotional manifestations among other emotions, the remaining data consisting in normal situations to ensure the occurence of a sufficient number of other emotional states or verbal interactions.

The fictional movie support has so far rarely been exploited for emotional computing studies[5]. On the one hand, fiction undoubtedly provides acted emotions and audio recordings effects which cannot always reflect a true picture of the situation. Furthermore, the audio and video channels are often remixed afterward and are recorded under better conditions than in real surveillance data. On the other hand, we are here working on data very different from laboratory data, which are taken out of context with clean recording conditions, and which have been largely studied in the past. The fiction provides recordings of emotional manifestations in their environmental noise. It offers a large scope of believable emotion portrayals. Emotions are expressed by skilled actors in interpersonal interactions. The large context defined by the

[5] We found only one paper (Amir and Cohen, 2007) which exploits dialog extracted from an animated film to study emotional speech.

movie script favours the identification of actors with characters and tends to stir genuine emotions. Besides, the emotional material is quite relevant from the application point of view. Various threat situations, speakers, and recording conditions are indeed illustrated. This diversity is required for surveillance applications. But the two major contributions of such a corpus are:

- the dynamic aspect of the emotions: the corpus illustrates the emotion evolution according to the situation in interpersonal interactions.
- the diversity of emotional manifestations: the fiction depicts a large variety of emotional manifestations which could be relevant for number of applications but which would be very difficult to collect in real life.

2.2 In situ description of the emotional content

We propose a task-dependent annotation strategy which aims both to define the emotional classes that will be considered by the system and to provide information to help understand system behaviours.

2.2.1 Annotation tools and strategy

The annotation scheme is defined via the XML formalism (eXtensive Mark-up Language) under ANVIL (Kipp, 2001) (Devillers et al., 2005a) which provides an appropriate interface for multimodal corpora annotation (see Figure 1).

The audio content description is carried out '*in situ*', which means in the context of the sequence and with the help of video support. It consists in the sequence description of both situational and emotional contents. The sequence is split into audio-based annotation units – the *segments*. These derive from the dialog and emotional structure of the interpersonal interactions. The segment corresponds to a speaker turn or a portion of speaker turn with an homogeneous emotional content, that is without abrupt emotional change, taking into account the following emotional descriptors (categorical and dimensional). This '*in situ*' description makes it possible to capture the evolution of the emotional manifestations occurring in a sequence and to study its correlation with the evolution of the situation.

2.2.2 Annotation tracks

The situation illustrated in the sequence is depicted by various contextual tracks:

- The *speaker track* provides the genre of the speaker and also its position in

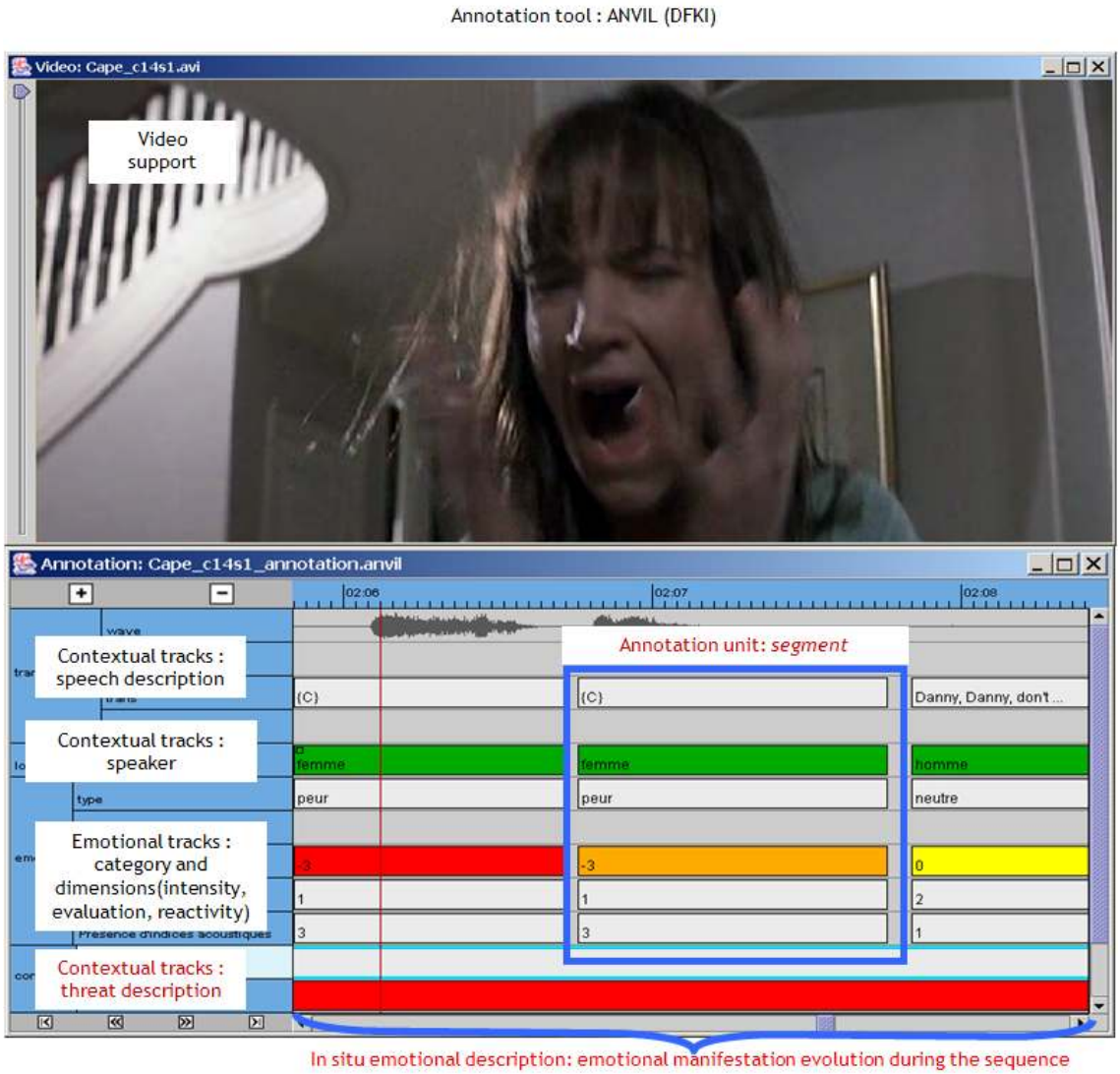


Fig. 1. *Annotation scheme under ANVIL*

the interaction (aggressor, victim or others).

- The *threat track* gives information about the degree of imminence of the threat (no threat, potential, latent, immediate or past threats) and its intensity. Besides, a categorization of threat types is proposed by answering the following step by step questions: If there is a threat, is it known by the victim(s)? Do the victims know the origin of the threat? Is the aggressor present in the sequence? Is he/she a familiar of the victims?
- The *speech track* stores the verbal and non-verbal (shouts, breathing) content of speech according to the LDC[6] transcription rules. The type of audio environment (music/noise) and the quality of speech are also detailed. The categories obtained via this annotation could be employed to test the robustness of the detection methods to environmental noise.

[6] Linguistic Data Consortium

Categorical and dimensional descriptors are used to describe the emotional manifestations at the segment level. Categorical descriptors provide a task-dependent description of the emotional content with various levels of accuracy. Indeed, it is especially difficult to accurately delimit the emotional categories (in terms of perceived classes for the annotation strategy and of acoustic models for the detection system see Section 1.1.2) when the data variability is high, as it is the case here. In order to limit the number of emotion classes, we have selected four major emotion classes: global class fear, other negative emotions, neutral, positive emotions. Global class fear corresponds to all fear-related emotional states and the neutral class corresponds to non-negative and non-positive emotional speech with a faint emotional activation, as defined in Devillers (2006) [7] . These broad emotional categories are specified by emotional subcategories which are chosen from a list of emotions occurring in abnormal situations. This list consists in both simple subcategories presented in Table 1 and mixed subcategories obtained by combining the simple subcategories (e.g. stress-anger).

Table 1
Emotional categories and subcategories.

Broad categories	Subcategories
fear	stress, terror, anxiety, worry, anguish, panic, distress, mixed subcategories
other negative emotions	anger, sadness, disgust, suffering, deception, contempt, shame, despair, cruelty, mixed subcategories
neutral	-
positive emotions	joy, relief, determination, pride, hope, gratitude, surprise, mixed subcategories

Dimensional descriptors are based on three abstract dimensions: evaluation, intensity and reactivity. They are quantified on discrete scales. Evaluation axis covers discrete values from wholly negative to wholly positive (-3,-2,-1,0,+1,+2,+3). The intensity and reactivity axes provide four levels from 0 to 3. The intensity dimension is a variant of the activation dimension defined in

[7] The concept of neutral emotion is ambiguous and needs to be clarified. The perception of "*neutral*" emotion is speaker-dependent and varies according to the "emotional intelligence" of the labellers (Clavel et al., 2006a). In this work, the "*neutral*" emotion corresponds to the cases where the judges could not perceive any emotion in the multimodal expression. Indeed, we are here focusing on the expressive aspect of the emotional phenomenon, that is one of the three aspects (cognitive, physiological, and expressive) currently accepted as composing the emotional phenomenon (Scherer, 1984). Harrigan et al. (2005) specify also this focus on the expressive aspect of emotion to define neutral attributes for their study.

psychological theories (Osgood et al., 1975) as the level of corporal excitation expressed by physiological reactions such as heartbeat increasing or transpiration. But we prefer to use the intensity dimension as we estimate it more suitable for the description of the oral emotional manifestations. For intensity and evaluation, the level 0 corresponds to neutral. The reactivity value indicates whether the speaker seems to be subjected to the situation (passive, level 0) or to react to it (active, level 3) and has been adapted to the application context from the most frequently used third dimension - named the control dimension (Russell, 1997). Besides this dimension is only used for emotional manifestations occurring during threats.

Abstract dimensions allow the specification of the broad emotional categories by combining the different levels of the scaled abstract dimensions. The perceptual salience of those descriptors and of the annotation unit was evaluated at the beginning of our work and as a preliminary step in validating the data acquisition and annotation strategy in Clavel et al. (2004).

2.2.3 Annotation task of labellers

The segmentation and the first annotation of the corpus were carried out by a native English labeller (Lab1). Two other French/English bilingual labellers (Lab2 and Lab3) independently annotated the emotional content of the pre-segmented sequences. It would be interesting to carry out further annotation exercises to strengthen the reliability but the annotation task is especially costly.

The contextual and video support of the sequence complicates the segment annotation and increases the annotation time. Indeed, the annotation of the emotional content in the pre-segmented sequences (7 hours) takes about 100 hours as the decision is taken by considering the context and the several channels (audio, video). But this support is crucial to strengthen the reliability of the annotations. The segmentation process is also very costly, since the complete segmentation and annotation task takes twice the time of the simple annotation of the pre-segmented sequences. Given the scale of this task we do not so far have a validation protocol for the segmentation step and for the other annotation tracks.

2.2.4 Evaluation of the reliability

When dealing with emotion computing, there are two main aspects to handle: the diversity of emotional manifestations and the subjectivity of emotion perception. We attempt to deal with the first aspect by considering various levels of accuracy in our annotation strategy. The second aspect is here unraveled by a comparative in-depth analysis of the annotations obtained by the three

labellers (Lab1, Lab2 and Lab3). The inter-labeller agreement is evaluated using traditional kappa statistics (Carletta, 1996) (Bakeman and Gottman, 1997) for the four emotional categories, and using Cronbach's alpha measure (Cronbach, 1951) for the three dimensions.

The kappa coefficient κ corresponds here to the agreement ratio taking into account the proportion of times that raters would agree by chance alone:

$$\kappa = \frac{\bar{p_o} - \bar{p_e}}{1 - \bar{p_e}}$$

where $\bar{p_o}$ is the observed agreement proportion and $\bar{p_e}$ the chance term. These two proportions are computed as follows: $\bar{p_0} = \frac{1}{N_{seg}} \sum_{i=1}^{N_{seg}} p_{seg_i}$ and $\bar{p_e} = \sum_{k=1}^{K} p_{cl_k}^2$. p_{cl_k} corresponds to the overall proportion of segments labelled with the class k, and the proportion p_{seg_i} corresponds to the measure of agreement on each segment i between the N_{ann} labellers. The kappa is at 0 when the agreement level corresponds to chance, and at 1 when the agreement is total.

Cronbach's alpha is another measure of inter-labeller reliability, more suitable than kappa for labels on a numerical scale. It is computed by the following formula:

$$\alpha = \frac{N_{ann}.\bar{r}}{1 + (N_{ann} - 1).\bar{r}}$$

where $\bar{r}$ is the average intercorrelation between the labellers. The higher the score, the more reliable the generated scale is. The widely-accepted social science cut-off is that alpha values at .70 or higher correspond to an acceptable reliability coefficient but lower thresholds are sometimes used in the literature (Nunnaly, 1978).

The Cronbach's alpha and the kappa statistics computed on the SAFE Corpus between the three labellers' annotations are presented in Table 2.

Table 2
Kappa score and Cronbach's alpha coefficient computed between the three labellers computed on the 5275 segments of the SAFE corpus

	Kappa	Cronbach
Categories (four categories)	0.49	.
Lab1 vs. Lab2	0.47	.
Lab1 vs. Lab3	0.54	.
Lab2 vs. Lab3	0.48	.
Intensity (four levels)	0.26	0.77
Evaluation (seven levels)	0.32	0.86
Reactivity (four levels)	0.14	0.55

The kappa score obtained for the agreement level of the four emotional categories between the three labellers is 0.49, which is an acceptable level of agreement for subjective phenomena such as emotions (Landis and Koch, 1977). Indeed, the use of global emotional categories allows us to obtain an acceptable level of agreement for the development of an automatic emotion recognition system. Moreover, this choice has been adopted by other studies: Douglas-Cowie et al. (2003), Shafran et al. (2003), Devillers et al. (2005b).

On the other hand, the kappas obtained for the three labellers are indeed much lower than for global categories. The kappa used here is the same as the one used for the categories and is dedicated to measure the level of strict agreement between the dimensional levels. The best kappa value is at 0.32 and is obtained for the evaluation axis from which the categories are derived. It shows that the level of strict agreement is poor and not sufficient to use the dimensions as distinct classes for the system. However the labellers' annotations according to the dimensions come out as correlated especially for intensity and evaluation, as illustrated by the high Cronbach's alpha values in Table 2. Each labeller seems to use his own reference scale on the dimension axis. However, this dimensional annotation provides interesting information to analyse the discrepancies between the labellers' annotations such as done in Clavel et al. (2006a).

For the system presented in this work, we make use of the annotation in global categories. For each category, we keep the data annotated as this category by the two labellers who have shown the highest disagreement on the entire corpus (the couple of labellers who has the lowest kappa). Segments for which these two labellers disagree are not considered for the system. This choice corresponds to a trade-off between the quantity and the reliability of the data considered for the training. Indeed, we did not choose to consider for the system the three annotations because the quantity of data where the three annotations converge is insufficient to build Gaussian mixture models. The intersection of two annotations allows us to obtain more data and the consideration of the two most divergent labellers (with the lowest kappa) ensures that on the data, where they agree, someone else would more probably also agree [8].

[8] Another solution to obtain a trade-off between the quantity and the reliability is to consider the segments where at least two of the evaluators agree. This configuration could be tested in a future work.

2.3 SAFE Corpus Content

2.3.1 Global content

Table 3 describes the SAFE corpus content in terms of sequences and segments. The segment duration depends on dialog interactions and on emotional variations in a speaker's turn. It follows that the segment duration is highly variable. The 5275 segments of the SAFE corpus represent 85% of the total duration of the corpus, and correspond to 6 hours of speech. The remaining 15% correspond to portions of recordings without speech, that is with silence or noise only.

Table 3
SAFE Corpus Content

Unit		minimum	maximum	mean
Segments	number of segments per sequence	1	53	13
	duration	40 msec.	80 msec.	4 msec.
	total number and duration	5275 - 6 h		
Sequences	duration	8 sec.	5 min.	1 min.
	total number and duration	400 - 7 h		

2.3.2 Sound environments

In most movies, recording conditions tend to mirror reality: speaker movements implying a natural variation in voice sound level are thus respected. However, the principal speaker will be audible more often in a fictional context. We can hypothesize that this is not systematically the case in real recording conditions. Overall, the sound environments are strongly dependent on the movie and may vary inside a movie and also inside a sequence. They depend on the type of situations depicted and their evolution.

The first diagram (Figure 2) indicates the segment distribution according to their sound environment (N= noise only, M = music only, N & M = noise and music, clean = without neither noise nor music) and the second diagram (Figure 3) according to speech quality (Q0=bad quality to Q3=good quality). 78% of the segments present an acceptable level of speech quality (Q2 or Q3) even though clean segments are not the majority part (21%) of the corpus. It shows that, despite the strong presence of noise or music, the speech quality is quite good. The corpus provides therefore a high number of exploitable segments. As presented further (see corpus SAFE_1 in Section 4.1), we select, for the system development and evaluation, segments with an acceptable level

of speech quality.

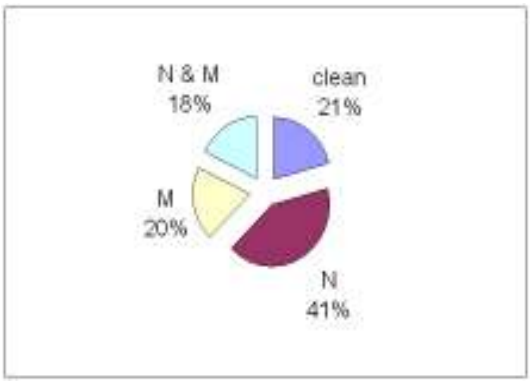


Fig. 2. *Segment division according to the sound environment.*

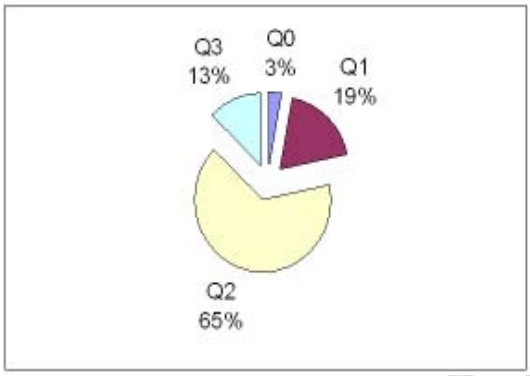


Fig. 3. *Segment division according to speech quality.*

2.3.3 Speakers

The surveillance application needs to cope with a high number of unknown speakers. With this in mind, the SAFE Corpus contains about 400 different speakers. The distribution of speech duration according to gender is as follows: 47% male speakers, 31% female speakers, 2% children. The remaining 20% of spoken duration consists in overlaps between speakers, including oral manifestations of the crowd (2%). We are aware of the need to process all the various types of spoken manifestations including overlaps for the ultimate application. However, the current work, given its exploratory character, does not take into account the 20% of the spoken duration consisting in overlaps (see corpus SAFE_1 in Section 4.1), as the acoustic modelling of fear is much harder in this case (e.g concurrent sources for pitch estimation).

2.3.4 Emotions

In this paper we emphasize the main features characterizing the emotional content of the SAFE corpus, that is the presence of extreme fear as illustrated by abstract dimension intensity and the relationship between the emotion label and context (threat). The emotional content is presented by considering the percentage of attributions for each label by the three labellers, so that the three annotations are taken into account [9] . The attribution percentage of the four emotional categories is thus the following: 32% for fear, 31% for other negative emotions, 29% for neutral, and 8% for positive emotions.

[9] We don't use the majority voting because there are more possible annotation choices than labellers. So there are segments where the three labellers may have three distinct annotations and which could therefore not be taken into account for the corpus description.

The attribution percentages of the various levels of the three dimensions is presented in the table 4. The reactivity is only evaluated on segments occurring during abnormal situations (71% of the segments) (see Section 2.2.2). In this context, the emotional manifestations are more majoritary associated with a low reactivity of the speaker to the threat. Very few segments are evaluated as positive on the evaluation axis (8% of positive emotions) and almost none of them is evaluated as level 3. Another specificity of our corpus consists in the presence of intense emotional manifestations: 50% of the segments are evaluated as level 2 or 3 on the intensity axis.

Table 4
SAFE Corpus Content: attribution percentage of emotional dimensions

levels / dimensions	-3	-2	-1	0	1	2	3
intensity	.	.	.	29%	21%	30%	20%
evaluation	9%	34%	20%	29%	5%	3%	0%
reactivity	.	.	.	6%	36%	19%	10%

Fear-type emotions are perceived as more intense than other emotions. 85% of fear segments are labelled as level 2 or 3 on the intensity scale while the major part of other emotions are labelled level 1. Besides, the presence of cries (139) seems to be associated with the presence of extreme fear.

2.3.5 Emotional manifestations and threat

The correlation of categorical descriptions of emotions with the threat provides a rich material to analyse the various emotional reactions to a situation. Figure 4 shows the distribution of each emotional category (fear, other negative emotions, neutral, positive emotions) as a function of the threat imminence. Fear is the major emotion during latent and immediate threats. By contrast fear is not very much in evidence during normal situations. Normal situations include a major part of neutral segments and also negative and positive emotions. Latent and past threats seem to cause a large part of other negative emotions than fear, which suggests that emotional reactions against a threat may be various.

Table 5 illustrates the segment distribution (%) according to each fear subcategory for each degree of imminence. The subcategory anxiety has almost never been selected by the labellers. Therefore we choose to merge the two subcategories worry and anxiety. The three last columns correspond to the most frequent mixed categories (see Section 2.2.2). Taken as a whole, the subcategories which are the most represented are thus anxiety-worry, panic and stress. During immediate threats, the major emotional sub-category is panic (31.8%). By contrast, normal situations include a major part of anxiety-worry

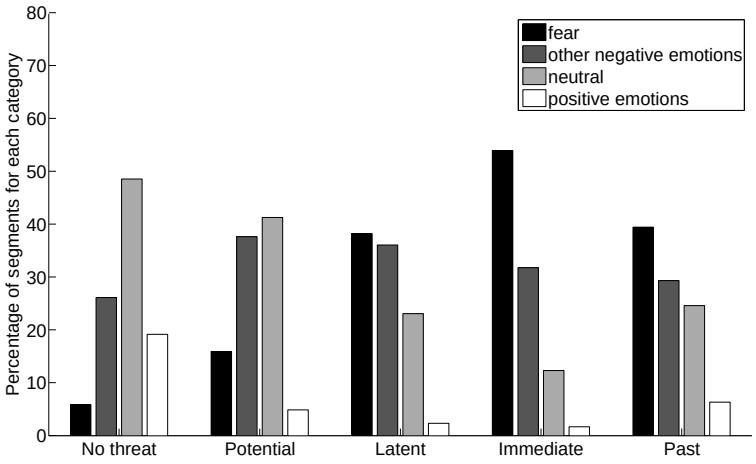


Fig. 4. *Segment distribution according to emotional categories for each degree of threat imminence*

but neither distress nor fear mixed with suffering, and very little terror and stress during normal situations. These sub-categories are almost exclusively present during immediate or latent threats. Otherwise, it is worth noting that fear frequently occurs mixed with other emotions such as anger (mostly), surprise, sadness and suffering.

During past threats, the subcategories anxiety-worry and panic are in the majority. The past threats which are depicted in the SAFE corpus are indeed threats occurring at the end of a sequence, and they occur just after immediate threats. It explains the presence of similar emotional manifestations as those occurring during immediate threats, yet with a higher proportion of anxiety caused by the threat.

Table 5
Segment distribution (%) according to each fear sub-category for each imminence degree (Pot. = potential, Lat. = latent, Imm. = immediate, Norm. = normal situation, anx. = anxiety, wor. = worry, ang. = anguish, distr. = distress, pan. = panic, terr. = terror)

threat \ emotion	anx. wor.	stress	ang.	distr.	pan.	terr.	fear-anger	fear-suffering	fear-surprise
Norm.	**64,1**	4,3	5,4	0,0	12,0	7,6	3,3	0,0	2,2
Pot.	**61,2**	4,1	6,1	0,0	10,2	6,1	10,2	0,0	2,0
Lat.	**50,9**	9,7	8,3	5,5	17,3	4,2	3,1	0,0	0,0
Imm.	25,2	13,7	6,7	6,0	**31,8**	8,1	3,5	2,7	0,8
Past	**42,0**	4,0	4,0	8,0	**22,0**	6,0	6,0	0,0	0,0

3 Carrying out an audio-based fear-type emotion recognition system

The fear-type emotion detection system focuses on differentiating fear class from neutral class. The audio stream has been manually pre-segmented into decision frames which correspond to the *segments* as defined in Section 2.2. The system is based on acoustic cues and focuses as a first step on classifying the predefined emotional segments.

3.1 Acoustic features extraction, normalization and selection

The emotional content is usually described in terms of global units such as the word, the syllable or the 'chunk' (Batliner et al., 2004), (Vidrascu and Devillers, 2005) by computing statistics. Alternatively, some studies use descriptions at the frame analysis level (Schuller et al., 2003). Here, we propose a new description approach which integrates various units of description that are at both the frame analysis level and the trajectory level. A trajectory gathers successive frames with the same voicing condition (see Figure 5). These two temporal description levels have the advantage of being automatically extracted.

Emotions in abnormal situations are accompanied by a strong body activity, such as running or tensing, which modifies the speech signal, in particular by increasing the proportion of unvoiced speech. Therefore some *segments* in the corpus do not contain a sufficient number of voiced frames. The information conveyed by the voiced content of the segment is therefore insufficient to deduce whether it is a fear segment or not. Such segments occur less frequently in everyday speech than in strong emotional speech. Here, 16% of the collected fear segments against 3% of the neutral segments contain less than 10% of voiced frames. The voiced model is not able to exploit those segments. Given the frequency of unvoiced portions and in order to handle this deficiency of the voiced model, a model of the emotional unvoiced content needs to be built. The studies which take the unvoiced portions into account consist of global temporal level descriptions (Schuller et al. (2004)), by computing for example the proportion of unvoiced portions in a 'chunk'. Our approach is original because it separately considers:

- the *voiced content* traditionally analysed and which corresponds to vowels or voiced consonants such as "b" or "d" and,
- the *unvoiced content* which is a generic term for both articulatory non voiced portions of the speech (for example obstruants) and portions of non-modal speech produced without voicing (for example creaky, breathy voice, mur-

mur).

The speech flow of each segment is divided into successive frames of 40 ms with a 30 ms overlap. The voicing strength of the frame is evaluated under Praat (Boersma and Weenink, 2005) by comparing the autocorrelation function to a threshold in order to divide the speech flow into voiced and unvoiced portions. Features are first computed frame by frame. In order to model the temporal evolution of the features, their derivatives and statistics (min, max, range, mean, standard deviation, kurtosis, skewness) are then computed at the trajectory level such as illustrated in Figure 5. Some features (the jitter, the shimmer and the unvoiced proportion) are computed at the *segment* level.

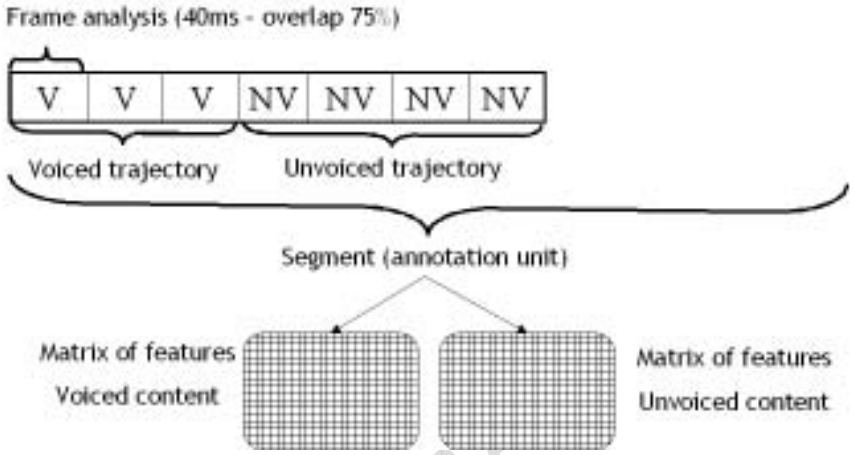


Fig. 5. *Feature extraction method which separately considers the voiced and unvoiced content and integrates various temporal levels of description*

The computed features allow us to characterize three types of acoustic content, and can be sorted into three feature groups:

- the *prosodic group* which includes features relating to pitch (F0), intensity contours, and the duration of the voiced trajectory, which are extracted with Praat. Pitch is computed using a robust algorithm for periodicity detection based on signal autocorrelation on each frame. Pitch and duration of the voiced trajectory are of course computed only for the voiced content;
- the *voice quality group* which includes the jitter (pitch modulation), the shimmer (amplitude modulation), the unvoiced rate (corresponding to the proportion of unvoiced frames in a given segment) and the harmonic to noise ratio (ratio of signal periodic part to non-harmonic part computed here using the algorithm developed in Yegnanarayana et al. (1998). The HNR allows us to characterize the noise contribution of speech during the vocal effort. The perceived noise is due to irregular oscillations of the vocal cords and to additive noise. The algorithm relies on the substitution degree for harmonics by noise.
- the *spectral and cepstral features group* consisting of the first two formants and their bandwidths, the Mel Frequency Cepstral Coefficients (MFCC),

the Bark band energy (BBE) and the spectral centroid (Cs).

A total of 534 features are thus calculated for the voiced content and 518 for the unvoiced content.

Acoustic features are not varying exclusively with the emotional content. They are also dependent on the speaker and on the phonetic content. It is typically the case for pitch-related features and for the first two formants. To handle this difficulty most of the studies use a speaker normalization for pitch-related features and a phoneme normalization for the first two formants. However the speaker normalization may be judged as inadequate to the surveillance, since the system needs to be speaker independent and has to cope with a high number of unknown speakers. The SAFE corpus provides about 400 different speakers for this purpose. The phoneme normalization is here also not performed as it relies on the use of a speech recognition tool in order to be able to align the transcription and the speech signal. The recording conditions of the speech signal in a surveillance application require to develop a text-independent emotion detection system which does not rely on a speech recognition tool. As a preliminary solution, we choose to use a min-max normalization which consists in the scaling of the features between -1 and 1. However, in future work, we plan to test more complex normalization techniques such as those used for speaker recognition (e.g. feature warping) and which might improve robustness to the mismatch of sound recordings and to noise.

The feature space is reduced by selecting the 40 most relevant features for a two class discrimination by using the Fisher selection algorithm (Duda and Hart, 1973) in two steps. A first selection is carried out on each feature group (prosodic, voice quality, and spectral) separately. One fifth of features is selected for each group providing a first feature set including about 100 features. The final feature set is then selected by applying the Fisher algorithm to the first feature set a second time . This method avoids having strong redundancies between the selected features by forcing the selection algorithm to select features from each group. The salience of the features is evaluated separately for the voiced and unvoiced contents. The Fisher selection algorithm relies on the computation of the Fisher Discriminant Ratio (FDR) of each feature i:

$$FDR_i = \frac{(\mu_{i,neutral} - \mu_{i,fear})^2}{\sigma^2_{i,neutral} + \sigma^2_{i,fear}}$$

where $\mu_{i,neutral}$ and $\mu_{i,fear}$ are class mean value of feature vector i for fear class and neutral class respectively and $\sigma^2_{i,neutral}$ and $\sigma^2_{i,fear}$ the variance values.

3.2 Machine learning and decision process

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 (Clavel et al., 2006b).

The classification is performed using the Gaussian Mixture Model (GMM) based approach which has been thoroughly benchmarked in the speech community. For each class C_q (*Fear*, *Neutral* and for each classifier (*Voiced*, *Unvoiced*) a probability density is computed and consists of a weighted linear combination of 8 Gaussian components $p_{m,q}$:

$$p(x/C_q) = \sum_{m=1}^{8} w_{m,q} p_{m,q}(x)$$

where $w_{m,q}$ are the weighted factors. Other model orders have been tested but led to worse results. The covariance matrix is diagonal which means that the models are trained by considering independently the data corresponding to each feature.

The parameters of the models (the weighted factors, the mean vector and the covariance matrix of each Gaussian component) are estimated using the traditional Expectation-Maximization algorithm (Dempster et al., 1977) with 10 iterations.

Classification is performed using the Maximum A Posteriori decision rule. For the voiced classifier, the A Posteriori Score (APS) of a segment associated with each class corresponds to the mean *a posteriori* log-probability and is computed by multiplying the probabilities obtained for each voiced analysis frame, giving for example for the voiced content:

$$\tilde{APS}_{voiced}(C_q) = \frac{\sum_{n=1}^{N_{f_{voiced}}} \log(p(x_n/C_q))}{N_{f_{voiced}}}$$

Depending on the proportion r of voiced frames ($r \in [0;1]$) in the segment, a weight (w) is assigned to the classifiers in order to obtain the final APS of the segment:

$$APS_{final} = (1-w) * APS_{voiced} + w * APS_{unvoiced}$$

The weight is dependent on the voiced rate ($r \in [0;1]$) of the segment according to the following function: $w = 1 - r^{\alpha}$. α varies from 0 (the results of unvoiced classifier are considered only when the segment does not contain any voiced frame) to $+\infty$ (only the results of unvoiced classifier are considered). The rate that the weight decreases as a function of the voiced rate is adjusted with α.

The segment is then classified according to the class (fear or neutral) that has the maximum *a posteriori* score:

$$q_0 = \arg \max_{1 \in [1:q]} \tilde{APS}_{final}(C_q)$$

4 Experimental validation and results

4.1 Experimental database and protocol

The SAFE corpus stands for the variability of spoken emotional manifestations in abnormal situations at several levels: in terms of speakers, sound etc. In order to restrict this variability given the exploratory character of this work, we focused here on the most prototypical emotional distinction, i.e. the fear vs. neutral discrimination. The following experiments and analysis are thus performed on a subcorpus containing only *good quality* segments labelled fear and neutral. The quality of the speech in the segments concerns the speech audibility and has been evaluated by the labellers (see Section 2.3). Remaining segments include various environment types (noise, music). Segments with overlaps between speakers have been discarded (see Section 2.3.3). Only segments, where the two human labellers who have obtained the lowest kappa value agree, are considered (see Section 2.2.4), i.e. a total of 994 *segments* (38% of fear segments and 62% of neutral segment). Table 6 shows the quantity of data corresponding to each class in terms of segment, trajectory and frame analysis. This subcorpus will be named *SAFE_1*.

Table 6
Experimental database SAFE_1 (seg. = segment, traj. = trajectories)

Classes	number of seg.	number of traj.	number of frames	duration
Fear	381	2891	113 385	19 min
Neutral	613	5417	181 615	30 min
Total	994	8308	295 000	49 min

The test protocol follows the *Leave One Movie Out* protocol: 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 then the omitted subset is used for the test. This protocol ensures that the speaker used for the test is not found in the training database [10].

[10] This is actually almost the case. Three speakers over the 400 speakers can be found in two films.

4.2 Global system behaviour

4.2.1 Selected features

It comes out from the feature selection step that pitch-related features are the most useful for the fear vs. neutral voiced classifier. With regard to voice quality features, both the jitter and the shimmer have been selected. The spectral centroid is also the most relevant spectral feature for the voiced content. As for the unvoiced content, spectral features and the Bark Band Energy in particular come out as the most useful.

Each classifier considers the features selected as the most relevant for the two-classes discrimination problem. Table 7 and Table 8 show the 40 selected features sorted by group for each content voiced or unvoiced. For the voiced content, the prosodic features are all selected after the second overall Fisher selection, which means that this feature group – especially the pitch related features – seems to be the most relevant for the fear-type emotions characterization. Voice quality features also seem to be relevant: both, jitter and shimmer have been selected. However the harmonic to noise ratio has not been selected. This may be explained by the presence of various environmental noise in our data which makes the HNR estimation more difficult. We should also keep in mind that the presence of music could bias the feature selection, such as the environmental noise. However, the segments which have been selected for the experiments contains also background music, but at a rather high signal (speech) to noise (music) ratio, so that this influence should not be detrimental.

The spectral and cepstral features which correspond to lower level features are preferred over the HNR. This feature group – initially in the greatest number – was the most represented in the final feature set. The most relevant spectral feature is the spectral centroid and cepstral features seem to be more relevant than features describing directly the spectral energy. Formants are also largely represented in the final feature set.

For the unvoiced content the Bark band energy-related features seem to be more relevant than cepstral features. The HNR has also been selected.

Overall, the selected features correspond to statistics computed at the trajectory level which seems to be suitable for emotional content characterization.

4.2.2 Voiced classifier vs. unvoiced classifier

Classification performance is evaluated by the equal error rate (EER). The EER corresponds to the error rate value occurring when the decision threshold

Table 7
List of the 40 selected features for the voiced content of SAFE_ 1 (Table 6). Nini1 = number of extracted features, Nini2 = number of features which are submitted to the second selection ($Nini2 = \lceil \frac{Nini1}{5} \rceil$), Nfinal = number of selected features at the end of the two successive selections, stdev = standard deviation, kurt = kurtosis, skew = skewness, d= derivative

Group	Nini2/Nini1	Selected features	Nfinal/Nini2
Prosodic	7/33	$meanF_0$, $minF_0$, F_0, $maxF_0$, $stdevdF_0$, $rangedF_0$, $rangeF_0$	7/7
Vocie quality	8/37	*Jitter*, *Shimmer*	2/8
Spectral	93/464	$meanC_s$, $minMFCC1$, $meanMFCC4$, $maxF_1$, $minMFCC4$, $mindF_1$, $mindF_2$, $meanMFCC1$, $rangedF_1$, $rangedF_2$, $rangeF_1$, $rangeF_2$, $MFCC4$, $MFCC1$, $stdevF_2$, $maxdF_1$, $maxdF_2$, $maxMFCC4$, $maxC_s$, $minMFCC3$, $skewBBE3$, $meanBBE3$, $maxF_2$, $stdevdMFCC11$, $stdevdF_2$, $kurtdF_1$, $minF_2$, $kurtF_1$, $rangeMFCC1$, $stdevdMFCC6$, $minMFCC6$	31/93

Table 8
List of the 40 selected features for the unvoiced content of SAFE_ 1 (Table 6).

Group	Nini2/Nini1	Selected features	Nfinal/Nini2
Prosodic	4/16	*rangeInt*	1/4
Voice quality	8/36	*tauxNonVoise*, *kurtdPAP*	2/8
Spectral	93/464	$rangeBBE6$, $rangeBBE7$, $rangeBBE10$, $stdevBBE6$, $stdevBBE7$, $rangeBBE8$, $rangeBBE5$, $stdevBBE10$, $rangeBBE11$, $rangeBBE9$, $stdevBBE8$, $rangeBBE12$, $maxMFCC3$, $stdevBBE5$, $stdevBBE9$, $rangeBBE4$, $rangeMFCC10$, $rangeMFCC12$, $stdevBBE11$, $minBBE10$, $rangeMFCC8$, $minBBE7$, $minBBE6$, $rangeBBE3$, $rangeMFCC6$, $minBBE8$, $rangedBBE6$, $minMFCC11$, $stdevBBE12$, $stdevBBE4$, $minBBE9$, $meanMFCC3$, $rangeMFCC11$, $rangedBBE5$, $maxBw_1$, $rangedBBE4$, $maxBw_2$	37/93

of the GMM classifier is set such that the recall will be approximately equal to the precision.

Figure 6 shows the EER for fear from neutral classification for various values of α. The voiced classifier is more efficient than the unvoiced one. The EER reaches 40% when the unvoiced classifier is used alone ($\alpha = \infty$). This worst case is equivalent to never considering the voiced content. However, the EER is at 32% when the voiced classifier is used in priority (the unvoiced classifier is used only when the segments are totally unvoiced, $\alpha = 0$). Best results ($EER = 29\%$) are obtained when the unvoiced classifier is considered with a weight decreasing quickly as the voiced rate increases ($\alpha = 0.1$).

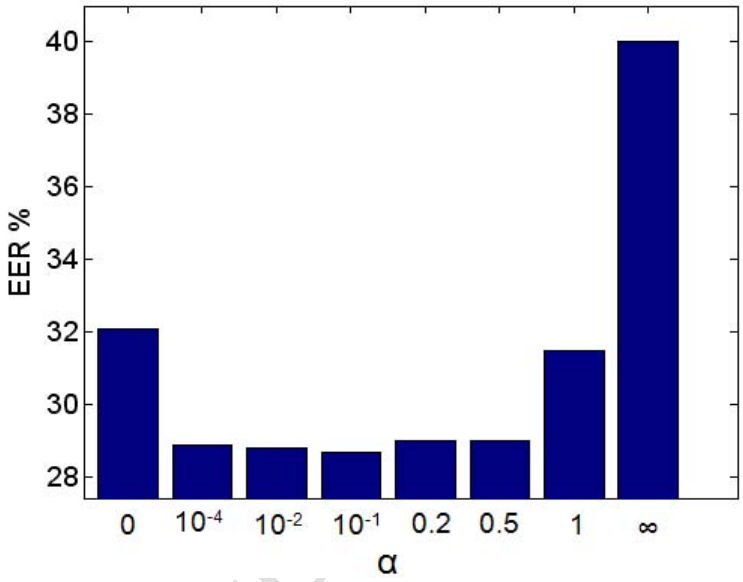


Fig. 6. *EER according to the weight ($w = 1 - r^{\alpha}$) of the unvoiced classifier against the voiced classifier obtained on SAFE_1 (Table 6) (confidence interval at 95%: radius $\leq$ 3%)*

The confusion matrix resulting from the fear vs. neutral classifier with the *alpha* parameter set at $\alpha = 0.1$ is presented in Table 9. It illustrates the confusions between the automatic labeling of the classifier and the manual labels provided by the labellers. We compute also the Mean Error Rate (MER) and the kappa (see Section 2.2.4) between human annotation and system classification for the performance evaluation. The kappa value at 0.53 corresponds here to the performance of the system taking into account the chance. This value integrates the unbalanced repartition of the data into the two classes and allows us to compare the system with chance (when $\kappa = 0$, the system is working such as chance).

The mean accuracy rate of the system is 71%. It corresponds to quite promising results given the diversity of fear manifestations illustrated in the SAFE Corpus (400 speakers, various emergence contexts and recording conditions).

Table 9
Confusion matrix, mean error rate (MER), equal error rate (EER) and κ for fear vs. neutral classification tested on SAFE_1 (Table 6) (confidence interval at 95%: radius $\leq$ 3%)

manual \ automatic	Neutral	Fear
Neutral	71%	29%
Fear	30%	70%
MER	29%	
EER	29%	
κ	0.53	

Otherwise, if one would expect deterioration of performance when trying to detect fear expressed in real context, performance could be improved by adapting the system to a specific sound environment and recording condition for a specific surveillance application.

We compute also the confusion matrix between the system outputs and each of the three labellers separately in Tables 10, 11, 12. The EER obtained are between 30% (when considering Lab3's annotation) and 35% (when considering Lab1's annotation) which means a 5% gap. It corresponds to similar results, given the confidence intervals.

Table 10
Confusion matrix for the fear vs. neutral classification system using Lab1's annotations as a reference (704 segments for neutral class and 631 segments for fear class, confidence interval at 95%: radius $\leq$ 4%)

Lab1 \ System	Neutral	Fear
Neutral	70%	30%
Fear	39%	61%
MER	34%	
EER	35%	
κ	0,48	

Table 11
Confusion matrix for the fear vs. neutral classification system using Lab2's annotations as a reference (1322 segments for neutral class and 518 segments for fear class, confidence interval at 95%: radius ≤ 4%)

Lab2 \ System	Neutral	Fear
Neutral	69%	31%
Fear	32%	68%
MER	32%	
EER	32%	
κ	0,45	

Table 12
Confusion matrix for the fear vs. neutral classification system using Lab3's annotations as a reference (352 segments for neutral class and 309 segments for fear class, confidence interval at 95%: radius ≤ 5%)

Lab3 \ System	Neutral	Fear
Neutral	72%	29%
Fear	31%	69%
MER	30%	
EER	30%	
κ	0.53	

4.2.3 System performance vs. human performance

A supplementary "blind" annotation based on the audio support only (i.e. by listening to the segments with no access to the contextual information conveyed by video and by the global content of the sequence) has been carried out by an additional labeller (LabSys) on SAFE_1. LabSys has to classify the segments into the categories *fear* or *neutral* with the same available information as the one provided to the system. We present in Table 13 the confusion matrix and the kappa score obtained by LabSys on SAFE_1. This table can be linked with Table 9 in order to compare the system performance with human performance.

The kappa obtained by the system is 0.53. It corresponds to a good performance compared to the value of 0.57 obtained by LabSys. However, the behaviors of LabSys and the system are quite different. LabSys is better to recognize *neutral* (99% of correct recognition against 71% for the system) and the system is better to recognize *fear* (70% of correct recognition against 64% for LabSys).

Table 13
Confusion matrix obtained by LabSys on SAFE_1

SAFE_1 \ LabSys	Neutral	Fear
Neutral	99%	1%
Fear	36%	64%
κ	0.57	

LabSys annotates more segments as neutral. Almost all the segments annotated *neutral* and 36% of those annotated *fear* in SAFE_1 are labelled *neutral* by LabSys. This shows that some fear cues are difficult to be perceived only with the audio channel.

4.3 Local system behaviour

Table 14 specifies the system behaviour on the various segments according to the threat during which they occur. With this aim, the emotional category annotations are correlated with the threat track annotations as presented in Clavel et al. (2007). Five fear subclasses are thus obtained:

- *NoThreat Fear*: fear occurring during normal situation, i.e. situation with no threat,
- *Latent Fear*: fear occurring during latent threats,
- *Potential Fear*: fear occurring during potential threats,
- *Immediate Fear*: fear occurring during immediate threats.
- *Past Fear*: fear occurring during past threats.

The reliability of the error rates *err* is evaluated by the 95% confidence interval (Bengio and Mariéthoz., 2004). The radius r of the confidence interval $I = [err - r; err + r]$ is computed according to the following formula:

$$r = 1,96\sqrt{\frac{err(1-err)}{N_{seg}}}$$

where N_{seg} is the number of segments used for the test.

The segment distribution of the fear class in the experimental database according to the type of the threat during which the segment occurs is presented in Table 14.

With regard to the fear recognition, we can see in Table 9 that 70% of the segments labelled fear are correctly recognized by the system. Best performances (78%) are obtained on *Immediate Fear* segments. By contrast, the recognition

Table 14
Proportion and recognition rate with confidence interval at 95% of fear segments according to the degree of imminence of the threat on SAFE_1 (Table 6)

		% of tested segments	recognition rate
fear	no threat	7%	61%±18%
	potential threat	4%	64%±24%
	latent threat	33%	60%±8%
	immediate threat	50%	78%±5%
	past threat	5%	71%±18%

rate falls on fear segments occurring during normal situation (61%±18%), potential (64%±24%) or latent (60%±8%) threats. Indeed, these last types of threats correspond to situations where the threat is not clearly present and where types of fear, such as anxiety or worry, frequently occur. In such segments, fear is less expressed at the acoustic level than in fear segments occuring during immediate or past threats, which explains the performance gap.

5 Conclusions and future work

The expectations in automatic emotion recognition/detection are ambitious. This research field is still emerging, and the emotional phenomenon remains especially complex to grasp. In this context our study corresponds to a preliminary work. So far, we have explored the different steps and strategies used in the development of a fear-type emotion recognition system dedicated to a given application, the audio-video surveillance. This innovative application has motivated us to take up new challenges in terms of emotional database and emotion recognition systems due to the specific class of the targeted emotions and the applicative constraints. Indeed, such an application implies to deal with heterogeneous data in noisy environments, which significantly makes more complex the classification task.

The first issue that we have addressed is the collection of recordings with emotional manifestations occurring in abnormal situations. Abnormal situations are especially rare and unpredictable and surveillance data are hardly accessible in order to protect the person privacy. Besides there is a lack of emotional databases (acted or real-life) which illustrate fear-type emotions in threat situations. The audiovisual corpus – the SAFE corpus – that we have built, contributes to handle this deficiency. We use a new material – fiction – to illustrate *in situ* emotional manifestations, including fear-type emotions (worry, terror, panic, etc.). More generally, the corpus contains recordings of both normal and abnormal situations and provides a large scope of contexts

and therefore a large scope of emotional manifestations. In this way, it forms an interesting support to study a high variety of emotional manifestations.

One of the lessons to be learned from our work, is that it is crucial to develop a detailed annotation scheme which allows us to better understand the variety of emotional manifestations and the associated system behaviour. One of our major contribution is to have defined an annotation strategy with various levels of accuracy which allows us both to better understand the variety of emotional manifestations and to provide computable emotional classes. Our annotation strategy has also the particularity to describe simultaneously the emotion evolution and the situation evolution. The annotation has been carried out by three labellers, and the three annotations have been confronted. This confrontation underlines the subjectivity of emotion perception and shows that our annotation strategy provides an acceptable level of agreement and constitutes a correct trade-off between genericity (data independent) and easiness of the labellers' task.

Another contribution which is worth mentioning is the dissociated description of the speech flow in terms of the voiced and unvoiced contents. This description has the advantage of considering the speech production peculiarities when the speaker is expressing strong emotions such as fear. We have extracted a large set of acoustic features and a selection of the most salient features has been performed using the Fisher selection algorithm. For the voiced content, the prosodic feature group – especially the pitch-related features – seems to be the most relevant for the fear-type emotion characterization, though voice quality features and lower level features, such as spectral and cepstral features, are also selected.

The fear vs. neutral classification achieves a mean accuracy rate of 71%. This is a quite promising result, given the diversity of fear manifestations illustrated in the SAFE Corpus (400 speakers, various emergence contexts and recording conditions). As the fear class gathers indeed a large scope of emotional manifestations which vary according to threats in particular, we have also studied the system behaviour on fear class according to the threat imminence. As expected, the best performance (78%) is obtained on fear segments occurring during immediate threats. In such segments, fear is indeed strongly expressed at the acoustic level with strong acoustic manifestations such as cries.

To sum up, the material used for our study is very complex. Given this complexity and the maturity of the field of emotion computing, we proceeded step by step by providing a first classification fear vs. neutral in order to overcome the complexity of the data. Indeed, it is important to deal with this complexity in terms of noisy speech and diversity of the data (speakers, situations) because it will be present in real audio surveillance data.

The discrimination between fear-type emotions and other emotions (e. g. positive and other negative emotions) will be one of the next steps of our study. Besides, it would be interesting to upgrade our system by modelling the evolution and the temporal context of the emotional manifestations. This dynamic aspect is already integrated into the annotation strategy and an analysis of emotional manifestations according to the threat imminence was performed. In a surveillance perspective, we would also like to change from the classification fear vs. neutral to the detection of fear-type emotions among other emotions.

Another challenge, which needs to be answered, is the processing of overlaps between speakers and of crowd emotional manifestations. This type of data are present in the SAFE corpus. They might provide acoustic cues characterizing group and crowd vocal manifestation during abnormal situations.

References

Abelin, A., Allwood, J., 2000. Cross linguistic interpretation of emotional prosody. In: Proc. of ISCA ITRW on Speech and Emotion. Belfast, pp. 110–113.

Amir, N., Cohen, R., 2007. Characterizing emotion in the soundtrack of an animated film: Credible or incredible? In: Proc. of Affective Computing and Intelligent Interaction. Lisbon, pp. 148–158.

Aubergé, V., Audibert, N., Rilliard, A., 2004. E-wiz: A trapper protocol for hunting the expressive speech corpora in lab. In: Proc. of LREC. Lisbon, pp. 179–182.

Bakeman, R., Gottman, J., 1997. Observing Interaction: an introduction to sequential analysis. Cambridge University Press.

Banse, R., Scherer, K., 1996. Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology 70 (3), 614–636.

Batliner, A., Fischer, K., Huber, R., Spilker, J., Nöth, E., 2003. How to find trouble in communication. Speech Communication 40 (1-2), 117–143.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., 2004. "you stupid ting box" - children interacting with the aibo robot: a cross-linguistic emotional speech corpus. In: Proc. of LREC. Lisbon, pp. 171–174.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Laskowski, K., Vogt, T., Devillers, L., Vidrascu, L., Amir, N., Kessous, L., Aharonson, V., 2006. Combining efforts for improving automatic classification of emotional user states. In: Proc. of Proc. IS-LTC. Ljubljana, pp. 240–245.

Bengio, S., Mariéthoz., J., 2004. A statistical significance test for person authentication. In: Proc. of Odyssey 2004: The Speaker and Language Recognition Workshop. Toledo.

Bänziger, T., Pirker, H., K.R., S., 2006. Gemep - geneva multimodal emotion

portrayals: A corpus for the study of multimodal emotional expressions. In: Proc. of LREC Workshop on Corpora for Research on Emotion and Affect. Genova, pp. 15–19.

Boersma, P., Weenink, D., 2005. Praat: doing phonetics by computer [computer program], from http://www.praat.org/. Tech. rep.

Breazeal, C., Aryananda, L., 2002. Recognizing affective intent in robot directed speech. Autonomous Robots 12 (1), 83–104.

Campbell, N., Mokhtari, P., 2003. Voice quality: the 4th. prosodic dimension. In: Proc. of International Congress on Phonetic Sciences. Barcelona, pp. 2417–2420.

Carletta, J., 1996. Assessing agreement on classification tasks: the kappa statistic. Computational Linguistics 22 (2), 249–254.

Clavel, C., Devillers, L., Richard, G., Vasilescu, I., Ehrette, T., 2007. Detection and analysis of abnormal situations through fear-type acoustic manifestations. In: Proc. of ICASSP. Honolulu, pp. 21–24.

Clavel, C., Ehrette, T., Richard, G., 2005. Events detection for an audio-based surveillance system. In: Proc. of ICME. Amsterdam, pp. 1306 – 1309.

Clavel, C., Vasilescu, I., Devillers, L., Ehrette, T., 2004. Fiction database for emotion detection in abnormal situations. In: Proc. of ICSLP. Jeju, pp. 2277–2280.

Clavel, C., Vasilescu, I., Devillers, L., Ehrette, T., Richard, G., 2006a. Fear-type emotions of the safe corpus: annotation issues. In: Proc. of LREC. Genoa, pp. 1099–1104.

Clavel, C., Vasilescu, I., Richard, G., Devillers, L., 2006b. Voiced and unvoiced content of fear-type emotions in the safe corpus. In: Proc. of Speech Prosody. PS6-10-222. Dresden.

Cowie, R., Cornelius, R., 2003. Describing the emotional states that are expressed in speech. Speech Communication 40 (1-2), 5–32.

Cronbach, L. J., 1951. Coefficient alpha and the internal structure of tests. Psychometrika 16, 297–334.

Damasio, A., 1994. Descartes'Error: Emotion, Reason and the Human Brain. Putnam publishing.

Darwin, C., 1872. The Expression of the Emotions in Man and Animals. Chicago University Press.

Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing emotion in speech. In: Proc. of ICSLP. Philadelphia, pp. 1970 – 1973.

Dempster, A., Laird, N., Rubin, D., 1977. Maximum likelihood from incomplete fata via the em algorithm. Journal of the Royal Statistical Society 39 (1), 1–38.

Devillers, L., December 2006. Les émotions dans les interactions homme-machine: perception, détection et génération. "Habilitation à Diriger des Recherches" Thesis, university Paris XI, Orsay.

Devillers, L., Abrilian, S., Martin, J.-C., 2005a. Representing real-life emotions in audiovisual data with non basic emotional patterns and context features. In: Proc. of ACII. Beijing, pp. 519–526.

Devillers, L., Vasilescu, I., 2003. Prosodic cues for emotion characterization in real-life spoken dialogs. In: Proc. of Eurospeech. Geneva, pp. 189–192.

Devillers, L., Vidrascu, L., 2007. Speaker characterization. Springer-Verlag, Ch. Emotion recognition.

Devillers, L., Vidrascu, L., Lamel, L., 2005b. Challenges in real-life emotion annotation and machine learning based detection. Journal of Neural Networks 18 (4), 407–422.

Douglas-Cowie, E., Campbell, N., Cowie, R., Roach, P., 2003. Emotional speech: Towards a new generation of databases. Speech Communication 40 (1-2), 33–60.

Duda, R., Hart, P. E., 1973. Pattern Classification and Scence Analysis. ser. Wiley-Interscience.

Ekman, P., 1999. Basic Emotions, handbook of cognition and emotion Edition. John Wiley, New York.

Ekman, P., Friesen, W. V., 1975. Unmasking the Face: A Guide to Recognizing Emotions from Facial Clues. Prentice-Hall.

Enos, F., Hirshberg, J., 2006. A framework for eliciting emotional speech: Capitalizing on the actor's process. In: Proc. of LREC Workshop on Corpora for Research on Emotion and Affect. Genova, pp. 6–10.

Fernandez, R., Picard, R. W., 2003. Modeling drivers' speech under stress. Speech Communication 40 (1-2), 145–159.

France, D., Shiavi, R., Silverman, S., Silverman, M., Wilkes, D., 2003. Acoustical properties of speech as indicators of depression and suicidal risks. IEEE Transactions on Biomedical Engeneering 47 (7), 829–837.

Harrigan, J. A., Rosenthal, R., Scherer, K. R., 2005. The New Handbook of Methods in Nonverbal Behavior Research. Oxford University Press (Eds), Oxford, UK.

Juslin, P., Laukka, P., 2003. Communication of emotions in vocal expression and music performance: different channels, same code? Psychological Bulletin 129 (5), 770–814.

Kienast, M., Sendlmeier, W.-F., 2000. Acoustical analysis of spectral and temporal changes in emotional speech. In: Proc. of ISCA ITRW on Speech and Emotion. Belfast, pp. 92–97.

Kipp, M., 2001. Anvil - a generic annotation tool for multimodal dialogue. In: Proc. of Eurospeech. Aalborg, pp. 1367–1370.

Kleiber, G., 1990. La sémantique du prototype, Catégories et sens lexical. PUF, Paris.

Kwon, O.-W., Chan, K., Hao, J., Lee, T.-W., 2003. Emotion recognition by speech signals. In: Proc. of Eurospeech. Geneva, pp. 125–128.

Landis, R., Koch, G., 1977. The measurement of an observer agreement for categorical data. Biometrics 33, 159–174.

Lee, C., Narayanan, S., Pieraccini, R., 2002. Classifying emotions in human-machine spoken dialogs. In: Proc. of ICME. Lausanne, pp. 737– 740.

Lee, C. M., Narayanan, S., Pieraccini, R., 1997. Facial emotion recognition using multi-modal information. Information, Communications and Signal

Processing 1, 347–401.
McGilloway, S., 1997. Negative symptoms and speech parameters in schizophrenia. Ph.D. thesis, Queen's University, Belfast.
Mozziconacci, S., 1998. Speech variability and emotion. Ph.D. thesis, Technical University Eidhoven.
Nunnaly, J., 1978. Psychometric theory. New York: McGraw-Hill.
Orthony, A., Turner, T., 1990. What's basic about basic emotion? Psychological Review 97, 315–331.
Osgood, C., mai, W. H., Miron, M., 1975. Cross-cultural Universals of Affective Meaning. University of Illinois Press, Urbana.
Oudeyer, P.-Y., 2003. The production and recognition of emotions in speech: features and algorithms. International Journal of Human Computer Interaction, special issue on Affective Computing 59 (1-2), 157–183.
Pelachaud, C., 2005. Multimodal expressive embodied conversational agent. In: Proc. of ACM Multimedia, Brave New Topics session. Singapour, pp. 683 – 689.
Picard, R., 1997. Affective Computing. MIT Press, Cambridge, MA.
Plutchik, R., 1984. A General Psychoevolutionary Theory. Vol. Approaches to Emotion. Erlbaum, Hillsdale, NJ.
Russell, J. A., 1997. How shall an emotion be called ? American Psychological Association,Washington, DC.
Scherer, K., 2003. Vocal communication of emotion : a review of research paradigms. Speech Communication 40 (1-2), 227–256.
Scherer, K., Schorr, A., Johnstone, T., 2001. Appraisal Processes in Emotion: Theory, Methods, Research. Oxford University Press.
Scherer, K. R., 1984. On the nature and function of emotion: A component process approach. Lawrence Erlbaum Associates, Publishers, Londres.
Scherer, U., Helfrich, H., Scherer, K. R., 1980. Internal push or external pull? Determinants of paralinguistic behavior. Oxford - New York: Pergamon.
Schroeder, M., Devillers, L., Karpouzis, K., Martin, J., Pelachaud, C., Peter, C., Pirker, H., Schuller, B., Tao, J., Wilson, I., 2007. What should a generic emotion markup language be able to represent? In: Proc. of ACII. Lisbon, pp. 440–451.
Schuller, B., Rigoll, G., Lang, M., 2003. Hidden markov model-based speech emotion recognition. In: Proc. of ICASSP. Hong Kong, pp. 1–4.
Schuller, B., Rigoll, G., Lang, M., 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proc. of ICASSP. Montreal, pp. 80–84.
Shafran, I., Riley, M., Mohri, M., 2003. Voice signatures. In: Proc. of ASRU Workshop. St Thomas, pp. 31 – 36.
Vacher, M., Istrate, D., Besacier, L., J.F.Serignat, Castelli, E., februar 2004. Sound detection and classification for medical telesurvey. In: Proc. of IASTED Biomedical Conference. Innsbruck, pp. 395–399.
van Bezooijen, R., 1984. Characteristics and recognizability of vocal expres-

sions of emotion. Foris Publications, Dordrecht.

Varadarajan, V., Hansen, J., Ayako, I., 2006. Ut-scope - a corpus for speech under cognitive/physical task stress and emotion. In: Proc. of LREC Workshop en Corpora for Research on Emotion and Affect. Genoa, pp. 72–75.

Vidrascu, L., Devillers, L., 2005. Detection of real-life emotions in call centers. In: Proc. of Eurospeech. Lisbon, pp. 1841–1844.

Wagner, J.; Vogt, T. E., 2007. A systematic comparison of different hmm designs for emotion recognition from acted and spontaneous speech. In: Proc. of ACII. Lisbon, pp. 114–125.

Whissel, C., 1989. The dictionary of affect in language. Emotion: Theory, Research, and Experience. New York: Academic Press.

Yacoub, S., Simske, S., Linke, X., Burns, J., 2003. Recognition of emotions in interactive voice response system. In: Proc. of Eurospeech. Geneva, pp. 729–732.

Yegnanarayana, B., d'Alessandro, C., Darsino, V., 1998. An iterative algorithm for decomposition of speech signals into periodic and aperiodic components. IEEE Transactions on Speech and Audio Processing 6 (1), 1–11.