The SAFE Corpus: illustrating extreme emotions in dynamic situations.

Chloé Clavel
Thales Research & Technology France
RD 128, 91767 Palaiseau Cedex
Chloe.clavel@thalesgroup.com

Ioana Vasilescu, Laurence Devillers
LIMSI-CNRS
BP 133, 91403 Orsay Cedex, France.
Ioana.vasilescu@limsi.fr, devil@limsi.fr

Gaël Richard
ENST-TSI
46 rue Barrault, 75634 Paris, Cedex 13, France
Gael.richard@enst.fr

Thibaut Ehrette, Célestin Sedogbo
Thales Research & Technology France
RD 128, 91767 Palaiseau Cedex
Thibaut.ehrette@thalesgroup.com

ABSTRACT
Existing real-life corpora illustrate everyday life contexts in which social emotions frequently occur. The type of emotional manifestations and the degree of intensity of such emotions are determined by politeness habits and cultural behaviours. This paper shows how the challenge of collecting extreme manifestations of emotion has been addressed with the acquisition of a corpus of fiction, the SAFE Corpus. The aimed application is civil safety and surveillance of public places in particular. A task-dependent annotation strategy is developed with both generic and specific descriptors. A description of the emotional content of the SAFE Corpus is provided. The corpus focuses on the illustration of extreme fear-type emotions in rich and varied contexts. Finally, a detection system of fear emotions based on acoustic cues has been developed to carry out an evaluation.

Keywords
Emotions in abnormal situation, civil safety, audiovisual corpus, annotation strategy, fear detection system.

INTRODUCTION

Emotions represent a complex research field with a large background in the communities of life and social sciences. In these contexts several theories have been proposed to define the concept "emotion" [5]. The emerging field of emotion analysis and detection in speech requires adapting these theories to the context of study and considering the variability inherent to the oral character.

In this paper, the aimed application is civil safety and surveillance of public places. Since current systems are mostly video-based, one of the main challenges is to use the audio content as complementary information to video to automatically detect an abnormal situation (situation during which the human life is in danger). The human oral communication in such situations is strongly based on the emotional channel. There is, as a consequence, a strong interest to automatically detect symptomatic emotions occurring in abnormal situations.

Studies dedicated to the analysis of emotion in speech commonly refer to a restricted number of emotions named "primary". The "Big Six" are, for example, frequently studied (fear, anger, sadness, joy, disgust and surprise). However emotions occurring, in everyday life contexts, are often a mix of several emotions or emotional manifestations which correspond more to the attitudes.

In this study, the targeted emotion is a primary emotion, namely *fear*. We are looking for fear-type emotions occurring in dynamic situations, during which the matter of survival is raised. In such situations the emotional manifestations correspond to primary manifestations of fear: they occur as a reaction to a threat. Their degree of intensity is particularly high and such emotions are rare in existing real-life corpora. Those corpora illustrate life contexts in which social emotions frequently occur. Abnormal situations are indeed rare and unpredictable and real-life recordings of abnormal situations are for the most confidential.

The lack of corpora illustrating strong emotions in real abnormal situations has encouraged us to build the SAFE Corpus (Situation Analysis in a Fictional and Emotional Corpus). The fiction provides 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.

The final goal is to develop a fear-type emotions detection system based on audio cues. The acoustic modeling of emotional speech obtained in such variable conditions is however more complex than for studies carried out on simulated emotions in laboratory conditions with a small number of speakers. The challenge is to control the variability of emotional manifestations by an appropriate annotation strategy. The annotation strategy has to be not only relevant for the application but also sufficiently *generic* to be exported to other corpora.

In the next section, the SAFE Corpus is presented. The third section is dedicated to the annotation scheme according to which we provide a description of the corpus content in the fourth section. Finally, the last section gives some

evaluation results of an automatic fear detection system using this corpus.

THE SAFE CORPUS

The SAFE Corpus consists of 400 audiovisual *sequences* from 8 seconds to 5 minutes extracted from a collection of 30 recent movies on DVD support. We focused for the sequence selection on the manifestation of emotional states in two contexts: normal vs. abnormal situations illustrated by individuals groups and/or crowds. Variability in terms of sequence duration depends on the way that situation is presented in the movie. Among the abnormal situations illustrated in the corpus we can mention natural damages such as fires, earthquakes, flood etc, physical or psychological threatening and aggression against human beings (kidnapping, hostages, etc.). A major contribution of such a corpus relies on the dynamic aspect of emotions: the corpus illustrates the emotion evolution according to the situation in interpersonal interactions. Besides, the fiction allows the collection of emotional data with their environmental noise. A total of 7 hours of recordings is thus collected in which speech represents 76% of the data. 71% of speech occurs in abnormal situations.

The movies make use mostly of American English (70% of the data). In the remaining movies, actors are portraying other English (British, Irish, Canadian) or foreign accents (French, Scandinavian, German).

The surveillance application implies to cope with a high number of unknown speakers. The SAFE Corpus provides about 400 different speakers in this purpose. The repartition of speech duration according to gender is as following: 47% male speakers, 32% female speakers, 1% child. The remaining 20% of spoken duration consists in overlaps between speakers, including oral manifestations of the crowd (2%).

ANNOTATING EMOTIONS IN DYNAMIC SITUATIONS

The annotation strategy takes into account the temporal aspect of the sequence. The emotion evolution is captured along the sequence by segmenting each sequence which provides a particular context into a basic annotation unit. This unit is called *segment* and corresponds to a speaker turn or a section of speaker turn portraying the same annotated emotion. The chosen descriptors are task-oriented. However the annotation strategy proposes also abstract descriptors which are more generic and can be exported to other applications. Annotation choices are helped by the video. A multimodal tool (ANVIL [7]) is used for the annotation.

It is especially difficult to delimitate accurate emotional categories (in terms of perceived classes for the annotation strategy and of acoustic models for the detection system) when the data illustrate a large degree of diversity. To overcome these challenges, the annotation strategy has been developed with the consideration of various levels of accuracy.

The segmentation and the annotation of the corpus were carried out by a first English native labeller. A second French/English bilingual labeller independently annotated the emotional content of the pre-segmented sequences. For these two labellers, the annotation of a given segment is influenced both by audio (acoustic and semantic content) and video information contained in the whole sequence. In order to evaluate the audio cues weight to detect a situation, which provokes fear, a supplementary "blind" annotation based on the audio support only have been done on a sub corpus [4]. This annotation is based on the listening to the segments in random order with no access to the contextual information conveyed by video and by the global content of the sequence.

Emotional Descriptors

The description of emotional substance is considered at the segment level and consists of two types of descriptors: dimensional and categorical. *Categorical descriptors* provide a task-dependant description of the emotional content with various level of genericity towards the corpus. We selected so far four major emotion classes: *global class fear*, *other negative emotions*, *neutral*, *positive emotions*. *Global class fear* corresponds to all fear-related emotional states. This last broad emotional category is completed by an oriented verbalization: the labeler has to precise the type of *fear* present in the segment by choosing a predefined (or not) sub-category (stress, terror, anxiety, etc.). *Dimensional descriptors* are based on the 3 abstract dimensions previously exploited in the literature [8]: activation, evaluation and control. The control dimension has been adapted here according to the application and renamed reactivity. The reactivity value indicates whether the speaker seems to be subjected to the situation (passive) or to react to it (active). Abstract dimensions are evaluated on discrete scales. The perceptual salience of those dimensional descriptors was evaluated in a former study [1]. Abstract dimensions allow specifying the broad emotional categories by combining the different levels of the scaled abstract dimensions.

Context Descriptors

The context of emotion emergence is described by a threat and a speaker track. The speaker track mentions the gender of the current speaker and its position in the interaction (*victim* or *aggressor*). The threat track provides the description of the threat intensity (4 levels scale) and of its incidence (*potential, latent, immediate, passed*).

Audio and Verbal Context Descriptors

Extracted sequences provide recordings made in variable conditions. Consequently, quality has a high variability inter- and intra-sequences. Labels defining audio quality of each segment are stored. We evaluate both the perceptual quality and we annotate the acoustic events which determine this quality. More precisely, the quality is estimated on a 4 steps scale: from *Q0* (inaudible speech or with too much sound effects) to *Q3* (perfectly intelligible speech and realistic sound recordings). The presence of noise and/or music is besides annotated according to 4

recording conditions: *Clean* (segments without noise and music), *Noise only*, *Music only* and *Noise and music*. The verbal and non verbal (cries, breathing, etc.) contents of the segments are also transcribed.

EMOTIONAL CONTENT

A total of 4073 speech segments with a duration ranging from 40 ms to 80s are obtained from the corpus sequences. The emotional content is presented in this section by considering the percentage of attributions for each label by the two labelers, so that the two annotations are taken into account. The inter-labeller agreement for the emotional content has been evaluated in another study [4]. The percentage of attributions of the four emotional categories is thus the following: 29% for *fear*, 30% for *other negative emotions*, 33% for *neutral*, and 8% for *positive emotions*. In this section we focus on the main features characterizing the emotional content: the presence of extreme fear as illustrated by abstract dimension intensity and the relationship between the emotion label and context (threat and environmental noise). A complete description of the corpus has been provided in a previous study [2].

Presence of Extreme Fear

The combination between categories and dimensions allows a best visibility of the types of manifestations contained by each category.

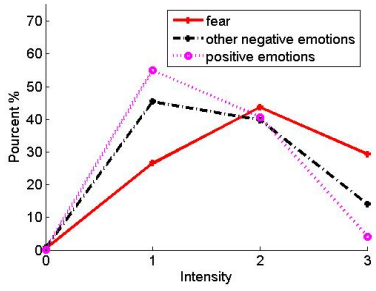

Figure 1. Distribution of each emotion category according to the intensity scale

Figure 1 shows the distribution of the attributions of each emotion category according to the first dimension, *intensity*. Fear-type emotions are perceived as more intense than other emotions. 73% 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.

Emotions in Dynamic Situations

The correlation of categorical descriptions of emotions with the threat provides a rich material to analyze the various emotional reactions to a situation. Table 1 shows the distribution of each emotional category *(fear, other negative emotions, neutral, positive emotions*) as a function of the threat incidence. 96% of *fear* segments occur in abnormal situations and 53% when the threat is immediate. We can notice the presence of other emotions than *fear* in abnormal situations. 61% of the segments labeled *other negative emotions* emerge during latent threats (30%) or immediate threats (31%).

	Abnormal situations/Threat				Normal/No Threat
	Potential	Latent	Immediate	Passed	None
Fear (29%)	4	32	**53**	7	**4**
Neg. (30%)	9	30	31	6	24
Neu. (33%)	7	25	14	7	47
Pos. (8%)	2	10	4	5	79

Tableau 1 Percentage of types of threat per emotional categories

Figure 2 shows the distribution of each intensity level inside the *fear* class according to the threat intensity. According to the table 1, 4% of segments labeled *fear* emerges in normal situations (No threat). It emerges from figure 2 that these segments (Threat Intensity = 0) corresponds to types of fear with a low level of emotional intensity such as anxiety or worry. 13% of *fear* segments labeled level 1 on the intensity scale occur in normal situation vs. 0.5% of *fear* segments labeled level 3. The segments labeled level 3 on the intensity scale correspond for the major part (57%) to threats with a higher level on the intensity scale. Besides *Fear* segments occurring when the threat is labeled with a low level of intensity are essentially labeled level 1 on the intensity scale

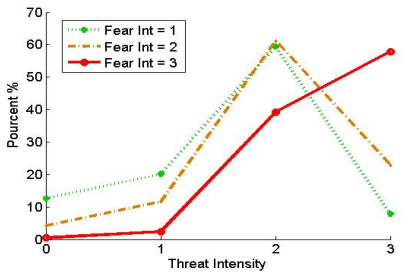

Figure 2. Emotional intensity of fear-type emotions according to threat intensity

Emotions in their Environmental Noise

Figure 3 shows that negative emotions correspond to noisier contexts than positive and neutral and contain more sound effects. In most movies, recording conditions tend to mirror reality: speaker movements implying natural variation in voice sound level are thus respected. However, the principal speaker will be more often audible in the fiction context. We can hypothesize that this is not systematically the case in real recording conditions. The categories obtained via

this annotation could be employed to the test of the robustness of detection methods to the environmental noise.

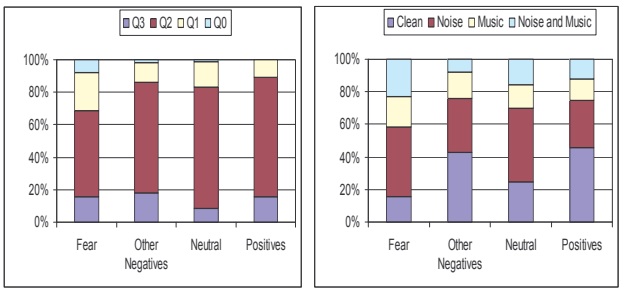

Figure 3. Emotional Categories vs. level of quality (left) and environment type (right)

EVALUATION RESULTS

We present here the first evaluation results of the fear-detection system developed on this corpus. The fear detection system is based on acoustic cues and focus as a first step on a fear vs. neutral classification of the emotional *segments*.

Experimental Database

The detection system is performed on a sub corpus containing only good quality (*Q2* or *Q3*) segments labeled *fear* and *neutral*. Overlaps have been avoided. Only segments where the two human coders agree are considered, i.e. a total of 986 segments (606 for *neutral* and 380 for *fear*). The gender repartition of the neutral segments is: 68% of male speakers, 30% of female speakers, 2% of child. The gender repartition of the fear segments is: 34% of male speakers, 62% of female speakers, 4% of child.

Fear versus neutral classifier

The classification system presented in details in [3] merges two classifiers, *the voiced classifier* and the *unvoiced classifier* which consider respectively the voiced portions and the unvoiced portions of the segment. The first step of the overall system aims at extracting prosodic and voice quality features and the second step at reducing the feature space using the fisher selection algorithm and the Principal Component Analysis. The third step consists in the training of the models of the two classes for each voicing condition (using Gaussian Mixture Models or GMM). The final step consists in the classification of each segment according to the two main classes (the *fear* class and the *neutral* class) merging the results of the two classifiers (*voiced* or *unvoiced*).

Test Protocol

The test protocol is the protocol *Leave One Movie Out*: the data are divided into 30 subsets; each subset contains all the segments of a movie. The protocol consists in training the model on 29 subsets, leaving the last subset for testing and in iterating the procedure for the 30 possible test subsets. This protocol ensures that the speaker used for the test is not found in the training database. Detection performances are evaluated by the equal error rate (EER). The EER corresponds to the error rate value occurring when the decision threshold of the GMM classifier is set so that the recall will be approximately equal to the precision. The corresponding chance performances are thus 50%.

Results

Best results (EER = 30.5%) are obtained when the unvoiced classifier is considered with a weight decreasing quickly when the voiced rate increases [3].

CONCLUSION & PERSPECTIVES

This paper is dedicated to the presentation of the SAFE Corpus. The challenge taken up by this corpus relies on the illustration of extreme fear-type emotions in threat dynamic contexts. A first *fear* vs. *neutral* detection system has been performed on this corpus. The EER obtained by this detection system is reaching 30.5%. Further work will be dedicated to the evaluation of a *fear* vs. *other emotions* detection system and to the building of specific acoustic models which takes into account the variability of *fear* manifestations.

REFERENCES

1. Clavel C., Vasilescu I, Devillers L., Ehrette T., Fiction database for emotion detection in abnormal situation, ICSLP 2004, Jeju.
2. Clavel C., Vasilescu I., Richard G. and Devillers L. Du corpus émotionnel au système de detection : le point de vue applicatif de la surveillance dans les lieux publics. Accepted for publication in the French Revue in Artificial Intelligence (RIA) (2006).
3. Clavel C., Vasilescu I., Richard G. and Devillers L. Voiced and Unvoiced content of fear-type emotions in the SAFE Corpus. Speech Prosody 2006, Dresden – to be published.
4. Clavel C., Vasilescu I., Devillers L., Ehrette T. and Richard G., SAFE Corpus: fear-type emotions detection for surveillance application, In Proc. International conference on Language Resources and Evaluation, Genoa, Italy, May 2006 – *to be published.*
5. Cowie R., Cornelius R., Describing the emotional states that are expressed in speech, Speech Communication – Speech and Emotion, volume 40, pages 5-32 (2003).
6. Douglas-Cowie, E., Campbell, N, Cowie R., Roach R., (2003), Emotional speech: Towards a new generation of databases, Speech Communication, vol 40, pages 33-60.
7. Kipp, M., (2001). Anvil a generic annotation tool for mul-timodal dialogue. Eurospeech.
8. Osgood, C., May, W.H., Miron M. S. (1975), Cross-cultural universals of affective meaning, University of Illinois Press, Urbana.