DETECTION AND ANALYSIS OF ABNORMAL SITUATIONS THROUGH FEAR-TYPE ACOUSTIC MANIFESTATIONS

C. Clavel[1,2,3], *L. Devillers*[3], *G. Richard* [2], *I. Vasilescu*[3], *T. Ehrette*[1]

[1]Thales Research and Technology France, RD 128, 91767 Palaiseau Cedex, France
[2]ENST-TSI, 37 rue Dareau, 75014 Paris, France
[3]LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France

ABSTRACT

Recent work on emotional speech processing has demonstrated the interest to consider the information conveyed by the emotional component in speech to enhance the understanding of human behaviors. But to date, there has been little integration of emotion detection systems in effective applications.

The present research focuses on the development of a fear-type emotions recognition system to detect and analyze abnormal situations for surveillance applications. The *Fear* vs. *Neutral* classification gets a mean accuracy rate at 70.3%. It corresponds to quite optimistic results given the diversity of fear manifestations illustrated in the data. More specific acoustic models are built inside the fear class by considering the context of emergence of the emotional manifestations, i.e. the type of the threat during which they occur, and which has a strong influence on fear acoustic manifestations. The potential use of these models for a threat type recognition system is also investigated. Such information about the situation can indeed be useful for surveillance systems.

Index Terms— emotional speech database, speaker independant fear recognition.

1. INTRODUCTION

These last years have seen an increase in interest for automatic surveillance systems[1]. Such systems are used as an assistance to humans which have to keep watching more than one place at a time. In such systems video cues have been largely used to detect abnormal situations : detection of abnormal objects, detection of crowd movements, etc. At the same time, audio events classification/detection is receiving a growing interest by the scientific community [2] [3].

It is especially the case in the context of audio retrieval and indexing applications but also in the context of multimedia event detection applications where audio can be used as a complementary source of information. However audio event detection has only begun to be used in some specific surveillance applications such as medical surveillance [4]. Audio cues, such as gun shots or screams [5] typically, may convey useful informations about the situation which can no longer be ignored in surveillance systems.

The goal of this paper is to develop an audio-based abnormal situations detection system in the context of civil safety. The targeted abnormal situations correspond to situations during which the human life is in danger (fire, physical or psychological attack, etc.). The human oral communication in such situations is strongly based on the emotional channel. Thus we choose to focus on the detection of emotional manifestations occurring in abnormal situations. More precisely the targeted emotions are fear-type emotions corresponding to symptomatic emotions occurring when the matter of survival is raised, including the different fear-related emotional states [6] from worry to panic.

Existing real-life corpora [7][8] illustrate everyday life contexts in which social emotions currently occur. The lack of corpora and studies dealing with strong emotions in real abnormal situations has encouraged us to build the SAFE Corpus (Situation Analysis in a Fictional and Emotional Corpus) [9] which consists of 7 hours of recordings extracted from fiction movies and which totals about 400 different speakers.

A fear-type emotions detection system based on acoustic cues has been developed using this corpus [10]. The targeted *Fear* class is a global class containing a high variability in terms of emotional representations. Fear manifestations are evolving according to the situation and according to the type of the threat (potential, latent, immediate or past) in particular. The basic idea of this paper is to model the various types of fear manifestations in order to derive information about the threat. With this purpose, the *Fear* class is divided into subclasses that are built according to the context, that is according to the type of the threat during which fear manifestations occur. Each type requires an appropriate intervention. There is therefore a strong interest to extract such information. A model for each subclass is built and we present here various classification strategies to the detection and the analysis of the threat.

In the next section, a description of the audio-based fear-type emotions detection system is provided. Then, in Section 3, the SAFE Corpus and the protocols used to evaluate the system are described. Finally, Section 4 presents the various classification strategies which have been tested and an analysis of the results.

2. THE FEAR-TYPE EMOTIONS DETECTION SYSTEM

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

The classification system merges two classifiers, the *voiced classifier* and the *unvoiced classifier* which consider respectively the voiced portions and the unvoiced portions of the segment [10]. The emotional manifestations conveyed by unvoiced speech portions needs indeed also to be modeled. Emotions in abnormal situations are indeed accompanied by a strong body activity, such as running or tens-

ing, which modifies the speech signal, by increasing the proportion of unvoiced speech in particular.

2.1. Feature extraction and selection

In this work, the emotional content is characterized by a large set of features including:

-*prosodic features* relating to pitch (F0), intensity contours and the duration of the voiced trajectory;

-*voice quality features* represented by 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;

-*spectral features* consisting in the first two formants and their bandwidths, the Mel Frequency Cepstral Coefficients (MFCC), the Bark band energy and the spectral centroid.

The acoustic content of each *segment* is represented with various levels of temporality. Features are computed every 10 ms on 40 ms-length frame analysis. In order to model the temporal evolution of the features, their derivatives and statistics (min, max, range, mean, standard deviation, kurtosis, skewness) are computed at more global temporal levels, corresponding for example to the voiced trajectory for pitch-related features or to the segment level for unvoiced rate. A total of 534 features are thus calculated. All the features are normalized so they are put on a single scale between -1 and 1. Silence frames are not considered and are automatically removed.

The feature space is reduced by selecting the 40 more relevant features for a two classes discrimination by using the Fisher selection algorithm [11] in two steps. A first selection is carried out on each feature family (prosodic, voice quality, and spectral) separately providing a first feature set. The final feature set is then selected by performing a second time the Fisher algorithm on the first feature set. This method ensures to avoid strong redundancies between the selected features by forcing the selection algorithm to select features from each family.

2.2. The training/classification steps

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 of each classifier (*Voiced Fear*, *Voiced Neutral*, *Unvoiced Fear* and *Unvoiced Neutral*) a probability density is computed and consists in 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 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 [12].

Classification is performed using the Maximum A Posteriori decision rule. For the voiced classifier, the A Posteriori Score (APS) of a segment associated to each class *Fear* or *Neutral* corresponds to the mean a posteriori log-probability and is computed by multiplying the probabilities obtained for each voiced analysis frame. The APS is computed in the same way for the unvoiced classifier. Depending on the proportion r of voiced frames ($r \in [0; 1]$) in the segment, a weight ($w = 1 - r^{\alpha}$) 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 parameter α has been previously set at $\alpha = 10^{-4}$ in [10] which means that the unvoiced classifier is considered with a weight decreasing quickly when the voiced rate increases.

3. THE SAFE CORPUS AND PROTOCOLS

3.1. Global Presentation

The SAFE Corpus consists of audio-visual sequences from 8s to 5min extracted from a collection of 30 recent movies in English language. Emotions are considered in their temporal context. We segmented each sequence that provides a particular context into a basic annotation unit, the *segment*, which has been defined in Section 2. 4724 segments of speech with a duration varying from 40ms to 80s are thus obtained from the 400 sequences of the corpus.

A *generic* annotation strategy was developed [9] and takes into account various aspects of the sequences content. The *emotional substance* is considered at the segment level and includes among other descriptors a description in four major emotion classes: *Fear*, *Other Negative Emotions*, *Neutral*, *Positive Emotions*. The *situational context* is described by a threat track and a speaker track (gender and identity of the speaker). The threat track describes the type of the threat (potential, latent, imminent, past) and its intensity. The *acoustic context* is described in terms of audio environment and speech quality.

Two labellers annotated the corpus. The segmentation and the annotation of the corpus were carried out by a first English native labeler. A second French/English bilingual labeler independently annotated the emotional content of the pre-segmented sequences. The inter-labeller agreement for the four emotional categories is evaluated thanks to the traditional kappa statistics [13]. The kappa score between the two labellers is at 0.47 which is an acceptable level of agreement for subjective phenomena such as emotions. We do not provide a validation protocol for the segmentation step because of the scale of this task.

3.2. Experimental Database

The following experiment and analysis are performed on a subcorpus containing only *good quality* segments labeled *Fear* and *Neutral*. The quality of the speech in the segments concerns the speech audibility and has been evaluated by the labelers. Remaining segments include various environment types (noise, music). Overlaps have been avoided. Only segments where the two human labelers agree are considered, i.e. a total of 994 *segments* (38% of *Fear* segments and 62% of *Neutral* segment). The emotional categories annotations are correlated with the threat track annotations. The segment repartition of the *Fear* class in the experimental database according to the type of the threat during which the segment occurs is stored in Table 1.

Fear				
No Threat	Potential	Latent	Immediate	Past
7.4%	3.7%	33.3%	50.1%	5.5%

Table 1. Segment repartition of the experimental database.

3.3. The Experimental Protocol

The test protocol follows the protocol *Leave One Movie Out* : 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, but using only the omitted subset for the test. This protocol ensures that the speaker used for the test is not found in the training database.

4. CLASSIFICATION STRATEGIES AND RESULTS

4.1. Abnormal situations recognition : *Fear*/*Neutral* classification

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

The confusion matrix resulting from the *Fear* vs. *Neutral* classifier is presented in Table 2. It illustrates the confusions between the automatic labeling of the classifier and the manual labels provided by the labelers. These results are obtained on the experimental database described in Section 3.2. The system behavior on the various segments according to the threat during which they occur is also detailed in this table. Due to the limited size of our database for potential or past threats (see Table 1), and to the situation proximity between potential (respectively past) threats and latent (respectively immediate) threats, results are considered separately on the *Fear* subclasses described in the Figure 1. The three subclasses corresponds to the +/- presence of the threat. One can notice that the past threats illustrated in the corpus correspond to contexts occurring just after the threat with as strong emotional manifestations as those occurring during immediate threats.

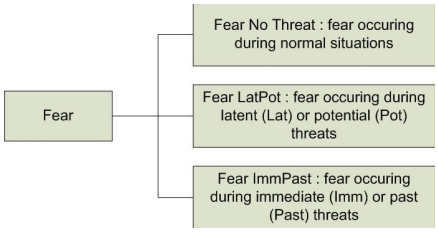

Fig. 1. *Fear subclasses*

manual \ automatic		Neutral		Fear	
Neutral		**71.3**		28.7	
Fear	NoThreat	29.7	39.3	**70.3**	60.7
	LatPot		39.0		61.0
	ImmPast		22.2		**77.8**
Mean Accuracy Rate		**70.8**			

Table 2. Confusion Matrix in percent of the *Fear* vs. *Neutral* classification system and system behavior according to the threat

The mean accuracy rate of the system is 70.8%. With regard to the fear recognition, 70.3% of the segments labelled *Fear* are correctly recognized by the system. Best performances (77.8%) are obtained on *Fear ImmPast* segments. Normal situations and latent or potential threats correspond to situations where the threat is not clearly present and where types of fear such as anxiety or worry 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 between *Fear NoThreat* (60.7%) or *Fear LatPot* (61.0%) segments and *Fear ImmPast* segments.

4.2. Threat type recognition: *Fear ImmPast* vs. *Fear LatPot*

In the previous framework, only two classes have been considered: *Fear* and *Neutral*. This system framework provides good performance when the threat is immediate or past. However when the threat is latent or potential, performance is decreasing.

The *Fear* class gathers indeed a large scope of emotional manifestations which are evolving according to the threat in particular. We propose here to build more specific acoustic models according to the type of the threat. The two subclasses *Fear LatPot* and *Fear ImmPast* represent the best trade off between independence (i.e.: acoustic proximities inside the subclasses) and future model quality (i.e.: sufficient number of members for training each subclass). The acoustic proximities between fear segments occurring during latent (respectively immediate) threats and those occurring during potential (respectively past) threats have been previously checked by performing a k-means unsupervised clustering on the segments as already done in [14]. Fear occurring during normal situation (*FearNoThreat*) will not be specifically modeled since it is not targeted in priority by the abnormal situation detection system.

The goal of this paragraph is to investigate the use of previous models to derive information about the threat. The previously described fear-type emotions recognition system indicates the presence of an abnormal situation. It is then interesting to provide supplementary information about the type of the threat by recognizing the various emotional manifestations inside the fear class. This information about the threat could indeed help humans to take the appropriate decision to limit the damage.

A classifier is associated to each fear subclasses. The fear recognition is now based on the merging of two classifiers : *Fear LatPot* vs. *Neutral* and *Fear ImmPast* vs. *Neutral*. Each classifier considers the features selected as the more relevant for the associated two-classes discrimination problem. Typically pitch related features are similarly selected by the two classifiers for the voiced content. That is not the case for example for formant and Bark Band energy related features which seem to be more relevant to the *Fear ImmPast* vs. *Neutral* discrimination. Inversely a higher number of MFCC related features is selected by the *Fear LatPot* vs. *Neutral* classifier.

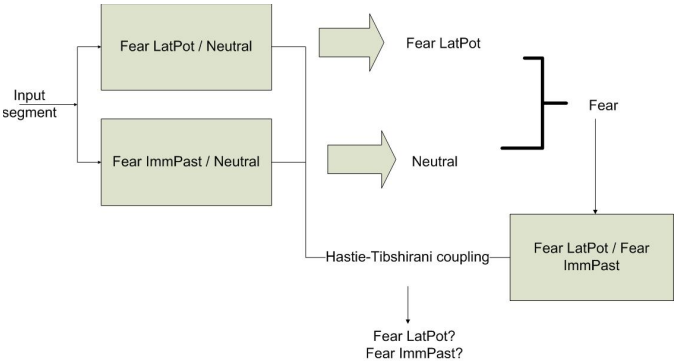

Fig. 2. *An example of the classification system running*

For each segment the a posteriori probability score corresponding to the two classes (*Fear LatPot*, *Fear ImmPast*) is computed and compared to the a posteriori probability score of the *Neutral* class. The classification *Fear* vs. *Neutral* is then performed using the following decision rule: *Fear* classification is decided if the segment is one of the following *Fear* classes (*Fear LatPot*, *Fear ImmPast*).

Segments which have been recognized as *Fear* are then submitted to a supplementary binary classifier *Fear LatPot* vs. *Fear ImmPast* as illustrated in Figure 2.

The *Fear LatPot* vs. *Fear ImmPast* classification is carried out using the Hastie-Tibshirani [15] approach to perform optimal coupling of the three classifiers as used in [3]. For a given test observation x_t, the likelihoods of each class *Fear LatPot*, $p(C_1|x_t)$ and *Fear ImmPast*, $p(C_2|x_t)$ are estimated by assuming the following model:

$$p_{i,j}(C_i|x_t) = \frac{p(C_i|x_t)}{p(C_i|x_t) + p(C_j|x_t)}$$

where $p_{i,j}(C_i|x_t)_{1\leq i,j\leq 3}$ correspond to the probability that x_t belongs to C_i considering the binary classifier $\{C_i, C_j\}$. The derived a posteriori scores of the two classes are then compared to perform the final classification.

The final results are stored in Table 3.

man. \ autom.	Neutral	Fear LatPot	Fear ImmPast
Neutral	60.9	20.9	18.2
Fear LatPot	34.0	**34.1** / 66.0	31.9
Fear ImmPast	13.2	26.9 / 86.8	**59.9**
M.A.R. Fear/Neutral		**69.3**	
M.A.R. LatPot/ImmPast/Neutral		**51.6**	

Table 3. Confusion Matrix in percent resulting from *Fear LatPot* vs. *Fear ImmPast* classification (M.A.R. = Mean Accuracy Rate)

59.9% of segments labelled *Fear ImmPast* are correctly recognized as immediate or past threat by the system and 34.1% of segments labelled *Fear LatPot* are also correctly recognized by the system as fear manifestations emerging in latent or potential threats. These first results could be improved by using the acoustic information as complementary information to visual cues and by considering the contextual informations. The information provided by the acoustic level may not be sufficient to differentiate subtle emotional states.

Even though the mean accuracy rate decreases from 70.8% to 69.3%, the global accuracy rate for the *Fear* class increases from 70.3% to 77.7% with this classifiers fusion framework. In particular this strategy enables us to enhance the performance for the detection of *Fear* for *Fear LatPot* segments: 66.0% of segments labelled *Fear* and occuring during potential or latent threats are correctly recognised as *Fear* by the system.

5. CONCLUSION

In this paper an abnormal situation detection system based on the recognition of fear-type emotions has been developed. The *Fear* vs. *Neutral* classification gets a mean accuracy rate at 70.3%. It corresponds to quite optimistic results given the diversity of fear manifestations illustrated in the SAFE Corpus (400 speakers, various emergence contexts and recording conditions) and the difficulty of the emotion recognition task. If one would expect deterioration of performance when trying to detect fear expressed in real context, performance could however be improved by adapting the system to a specific sound environment and recording condition for a specific surveillance application.

We have built specific models of fear manifestations according to the threat. These specific models have also led us to investigate the possibility to upgrade our system by providing a supplementary information about the threat. Future work will be dedicated to the correlation of information derived from fear acoustic manifestations with information derived from visual and contextual information to improve the robustness of the threat detection and analysis. Another prospect is the study of the correlation between more subtle fear manifestations such as anxiety or terror and the threat.

6. REFERENCES

[1] F. Dufaux and T. Ebrahimi, "Scrambling for video surveillance with privacy," in *Proc. of IEEE Workshop on Privacy Research In Vision*, New York, June 2006.

[2] R. Cai, L Lu, H.-J. Zhang, and L.-H. Cai, "Highlight sound effects detection in audio stream," 2003.

[3] S. Essid, G. Richard, and B. David, "Musical instrument recognition by pairwise classification strategies," *IEEE Transactions on Speech and Audio Processing.*, vol. 4, no. 1-2, pp. 1401–1412, July 2006.

[4] M. Vacher, D. Istrate, L. Besacier, J.F.Serignat, and E. Castelli, "Sound detection and classification for medical telesurvey," in *IASTED Biomedical Conference*, Innsbruck, Autriche, fevrier 2004, pp. 395–399.

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

[6] R. Cowie and R. Cornelius, "Describing the emotional states that are expressed in speech," *Speech Communication*, vol. 40, pp. 5–32, April 2003.

[7] V. Ververidis and C. Kotropoulos, "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162–1181, September 2006.

[8] L. Devillers, L. Vidrascu, and L. Lamel, "Challenges in real-life emotion annotation and machine learning based detection.," *Neural Networks*, vol. 18, no. 4, pp. 407–422, 2005.

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

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

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

[12] T. K Moon, "The expectation-maximization algorithm," *IEEE Signal Processing Magazine*, 1996.

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

[14] C. Clavel, I. Vasilescu, L. Devillers, T. Ehrette, and G. Richard, "Safe corpus: fear-type emotions detection for surveillance application," in *Language Resources and Evaluation Conference (LREC)*, Genoa, Italy, Mai 2006.

[15] Hastie T. and Tibshirani R., "Classification by pairwise coupling," Tech. Rep., Stanford University and university of Toronto, 1996.