Voiced and Unvoiced Content of fear-type emotions in the SAFE Corpus

Chloé Clavel[1,2,3], *Ioana Vasilescu*[2], *Gael Richard*[2], *Laurence Devillers*[3]

[1]Thales Research and Technology France, Domaine de Corbeville, 91404 Orsay Cedex, France
[2]ENST-TSI, 46 rue Barrault, 75634 Paris Cedex 13, France
[3]LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France
chloe.clavel@thalesgroup.com

Abstract

The present research focuses on the development of a *fear* detection system for surveillance applications based on acoustic cues. The emotional speech material used for this study comes from the previously collected SAFE Database (Situation Analysis in a Fictional and Emotional Database) which consists of audiovisual sequences extracted from movie fictions. We address here the question of a specific detection model based on unvoiced speech. In this purpose a set of features is considered for voiced and unvoiced speech. The salience of each feature is evaluated by computing the Fisher Discriminant Ratio for *fear* versus *neutral* discrimination. This study confirms that the voiced content and the prosodic features in particular are the most relevant. Finally the detection system merges information conveyed by both voiced and unvoiced acoustic content to enhance its performance. *fear* is recognized with 69.5% of success.

1. Introduction

Recent research on emotional speech reveals the need to go beyond the lexical and semantic levels of speech and to consider the emotional level which influences the semantic decoding of human interactions. In this way the emotional level shares in the improvement of speech processing systems. Furthermore, in dialog systems applications the determination of the speaker's emotional state aims at adapting the dialog strategy in order to provide a more relevant answer to the speaker's request [2], [6].

We address here the question of exploiting the speech emotional component to a new type of application, namely surveillance systems. Currently, surveillance systems dedicated to public places (bank, subway, airport etc.) tend to incorporate automatic video analysis to detect abnormal situations [12]. The goal is to use the audio content [5] as a complementary information to video. In particular we are interested here in the detection of symptomatic emotions occurring in abnormal situations. Abnormal situations are defined as contexts during which the human life is in danger. Extreme manifestations of negative emotions such as fear or other fear-related emotional states are thus expected to occur in these contexts.

Existing real-life corpora [7], illustrate everyday life contexts in which social emotions currently occur. The type of emotional manifestations and the degree of intensity of such emotions are determined by politeness habits and cultural behaviours. The emotions targeted by surveillance applications belong to the specific class of emotions emerging in abnormal situations. They occur indeed in dynamic situations, during which the matter of survival is raised. Abnormal situations are however rare and unpredictable and real-life recordings of such situations are for the most confidential. The SAFE Corpus (Situation Analysis in a Fictional and Emotional Corpus) has been built in order to provide an estimation of emotion acoustic particularities of fear-type emotions in abnormal situation. The fiction [3] 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. Emotions are emerging in interpersonal interactions in the heart of the action.

We address here the question of the detection of salient acoustic features which characterize fear-type emotions. Most of the studies focus on the voiced content of emotions which is known to convey relevant information about emotions. However emotions in abnormal situation are accompanied by a strong body activity, such as running or tensing, which might modify the speech signal, by for instance increasing the proportion of unvoiced speech.

In the following sections we describe the SAFE database (section 2) and the choice of acoustic features for fear-type emotion modelling (section 3). The salience of the acoustic features allowing to differentiate *neutral* from *fear* vocal manifestations in voiced and unvoiced portions of the speech is evaluated in section 4. Section 5 proposes a protocol allowing us to merge voiced and unvoiced information in a detection system, which discriminates *fear* from *neutral*.

2. The SAFE Corpus: description and annotation

The SAFE Corpus consists of audio-visual sequences from 8s to 5min extracted from a collection of 30 recent movies in English language. A total of 7 hours of recordings was collected in which spoken sequences represent 76% of the data. Emotions are considered in their temporal context. We segmented each sequence that provides a particular context into a basic annotation unit, the *segment*. It corresponds to a speaker turn or a section of speaker turn portraying the same annotated emotion. 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 with the view to be exported to other corpora and to a real life surveillance application [3], [4]. Various aspects of the sequences' content were taken into account: the emotional substance, the situational context (type of threat, speakers' gender and identity, location etc.) and the acoustic context (audio quality). Two labellers (1 English native, 1 bilingual French/English person) independently annotated the corpus.

The description of emotional substance is considered at the segment level and consists of two types of descriptors: dimensional and categorical. Categorical descriptors are employed for

the characterization of the emotional content of each segment. 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.

3. Feature extraction and pre-processing

The abnormal situation detection system focuses on differentiating *fear* from *neutral*. We present here the first crucial step of the detection system, namely modelling fear-type emotions. The goal of this section is to select acoustic features which allow us to optimally characterize fear-type emotions.

3.1. Prosodic and Voice Quality Features

The emotional content is usually characterize by classical prosodic features which help to describe the speech flow. They are perceived as stress, accentuation, rhythm and intonation and are thus relevant to characterize the speaker emotional state. Pitch-related features in particular play an important role in emotion recognition. In this paper *pitch* (F0) and *intensity* contours are extracted with Praat [10]. Pitch is computed using a robust algorithm for periodicity detection based on signal autocorrelation with 40 ms frame analysis. The last prosodic feature considered here is the *duration of the voiced trajectory*.

Emotional manifestations are not limited to prosodic variations [1] and the variations in terms of vocal effort are also carrying relevant information concerning the emotional state of the speaker. In this purpose we consider 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* (HNR) computed with Praat.

Voice quality is also characterized by spectral features such as *the first two formants and their bandwidths* computed by a LPC (Linear Prediction Coding) analysis. Perception-based spectral and cepstral features such as *Standard Mel Frequency Cepstral Coefficients* (MFCC), classically used in automatic speech recognition (ASR) and used more recently for emotion detection [11], *Bark band energy* and *spectral centroid* [8] are also considered.

The acoustic content of each segment is represented with various levels of temporality. Features are computed every 10 ms and stored in a matrix. In order to model the temporal evolution of each features, 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 157 features are thus calculated every 10 ms of each segment.

3.2. The question of normalization for surveillance application.

Some of the above features are varying not only with the emotional content. They are also dependant on the speaker and the phonetic content. It is typically the case for pitch-related features and 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 does not correspond to the surveillance application 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 in 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.

4. Features' salience in the voiced and unvoiced content

We evaluate here the salience of the previously described acoustic features to distinguish the two main emotional classes, *neutral* and *fear*. This analysis is performed on a subcorpus containing only *good quality* segments labelled *fear* and *neutral*. The quality of the speech in the segments has been evaluated by the coders. Overlaps have been avoided. Only segments where the two human coders agree are considered, i.e. a total of 1011 segments (665 for *neutral* and 345 for *fear*).

Some of the features can only be computed on voiced frames. However there are *segments* (see section 2) in the corpus which 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 15% of the collected *fear* segments against 2% of the *neutral* segments contain less than 10% of voiced frames. The voiced model is not able to exploit those segments. Given their frequency requiring a modelization and in order to handle this deficiency of the voiced model, a model of the emotional unvoiced content needs to be built. In this purpose the speech flow of each segment is divided with Praat into two types of vocal content:
- 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 fricatives) and portions of non modal speech produced without voicing (for example creaky, breathy voice, murmur).

The salience of the features is evaluated for the voiced and unvoiced contents separately. The Fisher Discriminant Ratio (FDR) of each feature i is computed for both contents :

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

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

4.1. Salient features of the voiced content

For this analysis only segments containing voiced frames are considered (329 *fear* segments and 664 neutral segments). The table 1 indicates the feature families which discriminate the best *fear* from *neutral*. A feature family corresponds to the feature, its derivative and its statistics. The corresponding FDR is also mentioned.

Results show that the voiced content is strongly represented by prosodic features and by pitch-related features in particular. Measures on pitch are strongly higher for *fear* than for *neutral*.

The jitter is among the most salient features for *fear* from *neutral* discrimination. It corresponds to a characterization of pitch modulation, i.e. to a *vibrato* in the voice which may be relevant to modelize cries. To confirm this assumption the FDR

Features family	FDR of the first selected feature of the family
Pitch	0.55 (mean)
Spectral Centroid	0.43 (skewness)
Jitter	0.30
Bandwith F1	0.17 (min)
F2	0.16 (standard deviation)
F1	0.13 (kurtosis of first derivative)
Bandwith F2	0.13 (min)

Table 1: Selected features for the voiced content

of the jitter is computed again by keeping only the 34 *fear* segments which contain cries. The result is relatively satisfying: the FDR is reaching 1.68 for segments containing cries versus *neutral* segments discrimination.

The selection of the first two formats (F1 and F2) raises the question of the dependence of these two features on the phonetic content. We decided to consider the formants analysis without any normalization by phonemes. We assume thus that the acoustic models of the formants do not depend on the phonetic content. This content would be thus similar for all emotion classes. In order to validate this assumption, the phonetic content of the *fear* class is compared to the phonetic content of *neutral* class. This comparison has been conducted by using a grapheme-to-phoneme transcriber on the verbal transcription. Figure 1 indicates the typical vowel repartition for the two classes and shows that the two phonetic contents are similar. The different behaviours of formants-related features according to the emotional classes are slightly influenced by the phonetic content and could be the effect of the emotional content.

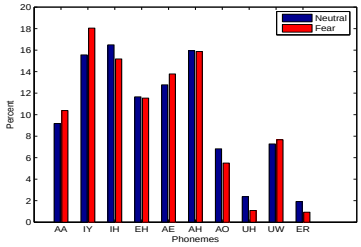

Figure 1: *Typical vowels repartition in the two emotional classes*

4.2. Salient features of the unvoiced content

For this analysis we consider the unvoiced portions of all the segments. The set of features considered in the case of unvoiced portions of the segments does not include the pitch and its statistics, the voiced trajectory duration and the jitter, since they can only be applied on voiced portions. The table 2 indicates the feature families selected as the most relevant for the distinction between the two classes for the unvoiced content. One may notice that the first selected features are less discriminant than those computed on voiced segments. Selected features are essentially spectral features represented by the formants and their bandwidths which are modelling here the vocal tract. MFCC are more relevant for unvoiced content with a FDR value three times higher than for the voiced discrimination ($FDR(MFCC)_{voiced} = 0.03$).

5. Voiced and unvoiced content in the detection system.

The goal is here to build a detection system based on *fear* versus *neutral* classification of the emotional *segments* (see section 2.1). 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 (see figure 2).

Features family	FDR of the first selected feature of the family
F2	0.15 (skewness)
Intensity	0.11 (kurtosis)
Unvoiced Rate	0.11
Mfcc	0.09
Bandwith F1	0.09 (standard deviation)
F1	0.09 (kurtosis of first derivative)

Table 2: Selected features for the unvoiced content

5.1. *Fear* versus *Neutral* GMM-based classifier

The first step of the overall system aims at reducing the feature space. The second 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).

- **Reduction of the feature space** : the feature space is reduced firstly by selecting the 60 more relevant features for the two classes discrimination (fisher selection algorithm) for each voicing condition and secondly by combining the different features to form a lower dimension vector (Principal Component analysis)

- **The training step by Gaussian Mixture Modelling** : for each class of each classifier (*voiced fear*, *voiced neutral*, *unvoiced fear* and *unvoiced neutral*) a Gaussian Mixture Model (GMM) is built. The parameters of the models are estimated using the traditional Expectation-Maximization algorithm [9] initialised by a basic binary splitting vector quantization algorithm.

- **The classification step** : classification is made using the Maximum A Posteriori (MAP) decision rule. For the voiced classifier, the mean a posteriori log-probability on the segment is computed for each class *fear* or *neutral* (by multiplying the probability obtained for each voiced time analysis frame). The mean a posteriori log-probability is computed in the same way for the unvoiced classifier. Depending on the proportion of voiced frames in the segment, a weight is attributed to the classifiers in order to obtain the final maximum a posteriori score of the segment. The segment is then classified according to the class (*fear* or *neutral*) that has the maximum a posteriori score. Silence windows are not considered and are automatically removed.

- **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. 30 training 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. Detection performances are evaluated by the equal error rate (EER). The EER corresponds to the error rate occurring when the decision threshold of GMM classifier is set so that the number of false rejections will be approximately equal to the number of false acceptances.

5.2. Experiments and results

The final maximum a posteriori score of a segment is computed for each class by summing the maximum a posteriori scores

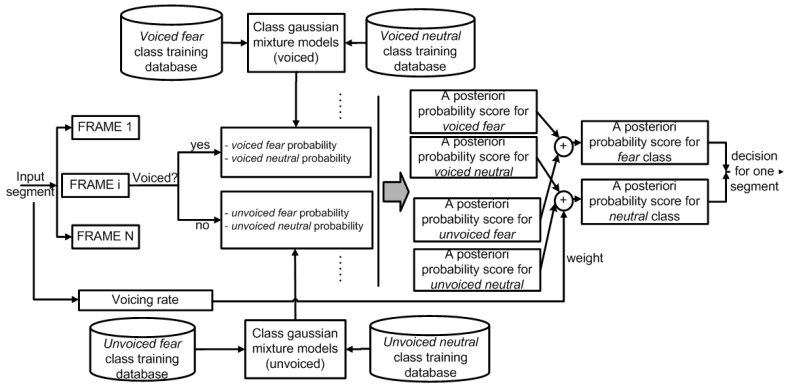

Figure 2: *Fear versus neutral classifier, merging voiced and unvoiced classifiers*

obtained for the voiced and unvoiced classifiers :

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

where w is the weight attributed to the unvoiced classifier against the voiced classifier. The weight is depending on the voiced rate ($r \in [0; 1]$) of the segment according to the following function :

$$w = 1 - r^{\alpha}$$

$alpha$ is varying from 0 (only the results of unvoiced classifier are considered) to $+\infty$ (the results of unvoiced classifier are considered only when the segment does not contain any voiced frame). The speed of the decreasing of the weight as a function of the voiced rate is adjusted with α. Figure 3 provides EER

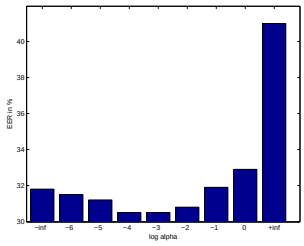

Figure 3: *EER according to the weight ($w = 1 - r^{\alpha}$) of the unvoiced classifier against the voiced classifier*

of *fear* from *neutral* detection for various values of α . The voiced classifier is more efficient than the unvoiced one. The EER is reaching 41.0% when the unvoiced classifier is used alone against 31.8% when the voiced classifier is used in priority (the unvoiced classifier is used only when the segments are totally unvoiced). Best results ($EER = 30.5\%$) are obtained when the unvoiced classifier is considered with a weight decreasing quickly when the voiced rate increases ($\alpha = 10^{-4}$).

6. Conclusions

In this paper a *fear* detection system based on both voiced and unvoiced emotional content has been built. *fear* is recognized with 69.5% of success. The detection system is based on specific models for each acoustic condition, voiced or unvoiced. The acoustic models are built by selecting the more relevant acoustic features to discriminate *fear* from *neutral*. The salience of each feature for the voiced and unvoiced contents is evaluated by computing the Fisher Discriminant Ratio. This evaluation highlights the discriminative power of the voiced content and of prosodic features in particular. Some of the features, such as jitter, seem to be particularly relevant to model segments containing the typical non verbal manifestation of *fear*, namely cries. Future work will be dedicated to the building of specific acoustic models of such extreme non verbal manifestations of *fear* with the view to develop a robust detection system of extreme manifestations.

7. References

[1] Aubergé, V.. Prosodie et émotion. *Actes des deuxièmes assises nationales du GdR I3*, pages 263-273, n French.

[2] Chateau, N.; Maffiolo, V.; Blouin, C. 2004. Analysis of emotional speech in voice mail messages: The influence of speaker's gender. *ICSLP*, Korea.

[3] Clavel, C.; Vasilescu, I.; Devillers, L.; Ehrette, T., 2004. Fiction Database for Emotion detection in Abnormal situation. *ICSLP*, Jeju: Korea.

[4] Clavel, C.; Vasilescu, I.; Richard, G.; Devillers, L., 2006. Du corpus émotionnel au système de détection : le point de vue applicatif de la surveillance dans les lieux publics. *Revue d'Intelligence Artificielle*, to be published in French.

[5] Clavel, C.; Ehrette, T.; Richard, G., 2005. Events detection for an audio-based surveillance system. *ICME*, Amsterdam: Netherlands.

[6] Devillers, L.; Vasilescu, I., 2004. Reliability of Lexical et Prosodic Cues in two Real-life Spoken Dialog Corpora. *LREC, Lisbon: Portugal.*

[7] Douglas Cowie, E.; Campbell, N.; Cowie, R.; Roach, P., 2003. Emotional speech: Towards a new generation of databases. Speech Communication.

[8] Ehrette, T.; Chateau, N.; d'Allessandro, C.; Maffiolo V., 2003. Predicting the Perceptive Judgment of Voices in a Telecom Context: Selection of Acoustic Parameters. Eurospeech, Geneva: Switzerland.

[9] Moon, T. K., 1996. The Expectation-Maximization algorithm, *IEEE Signal Processing Magazine.*

[10] http://www.praat.org

[11] Shafran, I.; Riley, M.; Mohri, M., 2003. Voice Signatures. *Automatic Speech Recognition and Understanding Workshop*.

[12] http://www.inria.fr/MULTIMEDIA/Videotheque/0-Fiches-Videos/517-fra.html