Fear-type emotions of the SAFE Corpus: annotation issues

Chloé Clavel[1,2,3], Ioana Vasilescu [2], Laurence Devillers[2], Thibaut Ehrette[1] and Gaël Richard[3].

(1) Thales Research & Technology France, RD 128, 91767 Palaiseau Cedex

(2) LIMSI-CNRS, BP 133, 91403 Orsay Cedex, France.

(3) ENST-TSI, 46 rue Barrault, 75634 Paris, Cedex 13, France

E-mail: chloe.clavel@thalesgroup.com

Abstract

The present research focuses on annotation issues in the context of the acoustic detection of fear-type emotions for surveillance applications. 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 audio-visual sequences extracted from movie fictions. A generic annotation scheme was developed to annotate the various emotional manifestations contained in the corpus. The annotation was carried out by two labellers and the two annotations strategies are confronted. It emerges that the borderline between *emotion* and *neutral* vary according to the labeller. An acoustic validation by a third labeller allows at analysing the two strategies. Two human strategies are then observed: a first one, context-oriented which mixes audio and contextual (video) information in emotion categorization; and a second one, based mainly on audio information. The k-means clustering confirms the role of audio cues in human annotation strategies. It particularly helps in evaluating those strategies from the point of view of a detection system based on audio cues.

1. Introduction

The emotional information conveyed by speech has been ignored for a long time and speech and language studies have mostly focused on the explicit message provided by the lexical level. Recently the trend has been changed and a fad for the emotional phenomenon has emerged (Cowie et al, 2001). The emotional content influences the semantic decoding of human interactions and can in this way share in the improvement of speech processing systems. Furthermore, in dialog systems applications the identification 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 (Lee et al, 2001, Devillers et al., 2003).

In this study, we are interested in audio-video surveillance applications. 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.

This growing interest for research on emotions has raised the question of collecting and annotating corpora of emotional speech. The emotional phenomenon is subtler, rarer and more subjective than the phenomena previously studied in dialog (Craggs, 2004).

Obviously, the performances of an automatic emotion detection system strongly depend on the quality of the emotional material used to build the acoustic models. The challenge is to delimitate accurate emotional categories both in terms of perceived classes for the annotation strategy and in terms of acoustic models for the detection system.

Collecting appropriate emotional data in the context of surveillance applications is a difficult task. The emotions targeted by surveillance applications belong indeed to the specific class of emotions emerging in abnormal situations. They occur 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. Existing real-life corpora (Douglas-Cowie, 2003), 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.

However, 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). A *fear-type* emotions detection system based on acoustic cues has been developed using this corpus (Clavel et al, 2006). The targeted *fear* class is a global class containing a high variability in terms of emotional representations. In particular, *fear* is largely represented in terms of emotional intensity. However, the annotation of this emotional intensity is quite subjective and depends on the labeller annotation strategy. It is shown in this paper that the apparent discrepancies between the two labellers can be, to a large extent, explained by the different strategies adopted.

The paper is organised as follows. In the next section, the SAFE Corpus is briefly described. Then, in section 3, a comparative in-depth analysis of the annotations of the two labellers is provided including an experiment on segments that were re-annotated using only the audio data (without access to the corresponding video data). In section 4, the potential influence of the various human annotation strategies on the automatic detection system performances is discussed.

2. The SAFE Corpus

2.1 Global Content

The SAFE Corpus consists of 400 audio-visual ***sequences*** from 8s to 5min extracted from a collection of

30 recent movies in English. 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. Emotions are emerging in interpersonal interactions in the heart of the action.
Even if the audio and video acquisitions are conducted under better conditions than it would be in a real audio and video surveillance, the fiction allows the collection of emotional data with their environmental noise.
A total of 7 hours of recordings was collected in which speech represents 76% of the data. Each sequence corresponds to a particular situation, normal or abnormal. Emotions are considered in the temporal context of the *sequence*.

2.2. Annotation Scheme

A ***generic*** annotation scheme was developed with the view to be exported to other corpora and to a real life surveillance application (Clavel et al, 2004, 2006). Various aspects of the sequence content were taken into account: ***the emotional substance, the situational context*** (type of threat, speaker gender and identity, location etc.) and the acoustic context (audio quality). Video was used as help for the annotation via the use of a tool for multimodal annotation ANVIL (Kipp, 2001). The segmentation and the annotation of the corpus were carried out by a first English native labeller.
Each sequence was segmented into a basic annotation unit, the ***segment***. It corresponds to a speaker turn or a section of a 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.
The description of emotional substance is considered at the segment level and is broken into two types of descriptors: dimensional and categorical. ***Dimensional descriptors*** are based on the three abstract dimensions previously exploited in the literature (Osgood, 1975): activation, evaluation and control. The control dimension has been here adapted according to the application and renamed reactivity. Abstract dimensions are evaluated on discrete scales. The perceptual salience of those descriptors was evaluated in a former study (Clavel et al, 2004). ***Categorical descriptors*** are also employed for the characterisation of the emotional content of each segment. Four major emotion classes have been selected: global class fear, other negative emotions, neutral, positive emotions. Global class fear corresponds to all fear-related emotional states.
A second labeller independently annotated the emotional content of the pre-segmented sequences. From now on, the English native labeller and the bilingual labeller are respectively named ***Lab1*** and ***Lab2***.

3. Human annotation strategies

In this section, the level of agreement between the two labellers, *Lab1* and *Lab2*, in emotional annotations is evaluated and the two human annotation strategies are confronted.

3.1. Categorization Process

The inter-labeller agreement for the four emotional categories is evaluated thanks to the traditional kappa statistics (Seigal, 1988). The kappa score between the two labellers is at 0.46 which is an acceptable level of agreement for subjective phenomena such as emotions (Craggs, 2004). Table 1 shows the repartition of the segments among the emotional categories according to the labeller. It emerges that *Lab1* evaluates more segments (75%) as non-neutral, than *Lab2* (58%).

	Fear	Other negative emotions	Neutral	Positive emotions
Lab1	32%	35%	**25%**	8%
Lab 2	27%	24%	**42%**	7%

Table 1: Emotional categories repartition according to each labeller.

In order to highlight where the disagreements are located, figure 1 illustrates the confusions between the two labellers for each emotional category. For example the histogram associated with *fear* on the x-axis takes into account the segments labelled *fear* by *Lab2*. It shows the distribution of the labels selected by *Lab1* for these segments. *Lab1* agrees on 78% of the segments labelled fear by *Lab2*. On the other hand 53 % of the segments labelled *neutral* by *Lab2* are labelled *emotion* (*fear*, *other negative emotions* or *positive emotions)* by *Lab1*. Therefore the major cause of disagreement between the two labellers relies on the *emotion* versus *neutral* categorisation. The second cause of disagreement is due to confusion between *fear* and *other negative emotions* (15% of the segments labelled *fear* by the *Lab2* are labelled *other negative emotions* by *Lab1* and 22% of the segments labelled *other negative emotions* are labelled as *fear*). We can also notice a few confusions between positive emotions and negative emotions which are due to mixed emotions, difficult to annotate.

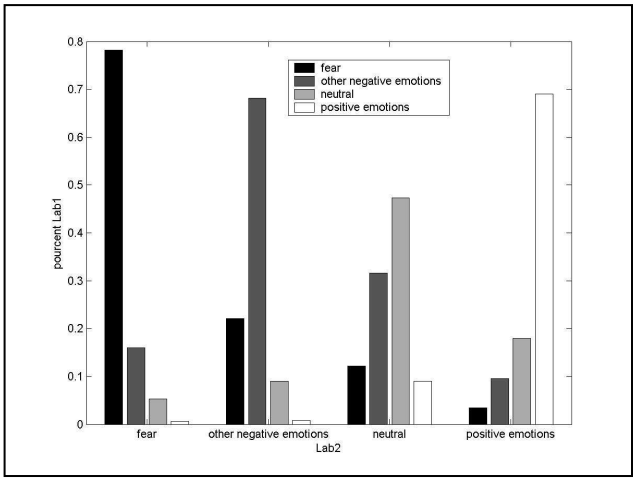

Figure 1: Confusion histogram for emotional categories

3.2. Intensity scale labelling

The ***borderline*** between ***emotion*** and ***neutral state*** is not the same for the two labellers. This finding encourages us to correlate emotion categories with their dimensional descriptions and with the intensity dimension in particular.

The kappa is also computed for the intensity scale. The kappa score between the two labellers is at 0.24. This score is quite low. However it measures the level of agreement between the four levels (0, 1, 2, 3) of the intensity scale without considering the level proximity. The Cronbach's alpha measure (Cronbach, 1951) is another measure of inter-labeller reliability, more suitable than kappa for labels on a numerical scale such as those used for the intensity axis. The Cronbach's alpha score is here at 0.82. This score shows that, even though the degree of agreement on the four intensity levels intensity between the two labellers is quite low, the two labellers annotation strategies seems to be correlated.

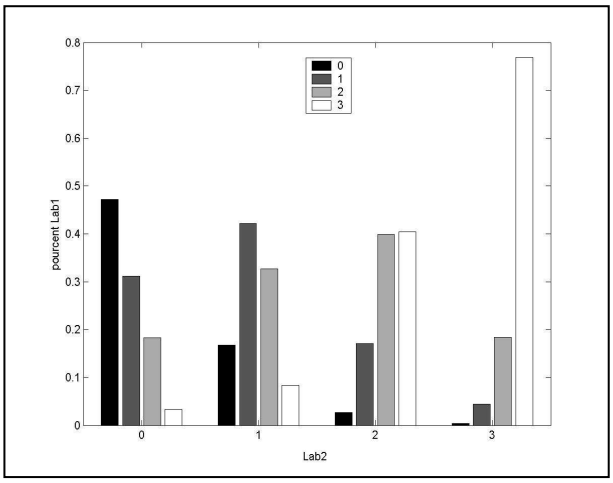

Figure 2: Confusion histogram for the intensity scale

In order to locate the disagreement, the confusion histogram for the intensity scale is represented on figure 2. As expected, the segments labelled *neutral* (intensity level = 0) by *Lab2* and *emotion* by *Lab1* are for the major part labelled level 1 on the intensity scale. Globally *Lab1* evaluates the segments as more intense with one level higher than *Lab2*. 77% of the segments labelled level 3 by *Lab2* are evaluated with the same level by *Lab1*. On the other hand 42% of the segments labelled 2 by *Lab2* are labelled 3 by *Lab1*.

3.3. "Blind" annotation

The annotation of a given segment is influenced both by audio (acoustic and semantic content) and video information contained in the whole sequence. The SAFE Corpus is indeed audio-visual and the audio annotation under ANVIL is done using the video stream. This choice is driven by the final perspective of this research. The audio module is aimed to be inserted in a multimodal surveillance system. However the goal of the current research is to develop an emotion detection system based on acoustic cues. The acoustic detection system previously developed (Clavel et al, 2006) relies on a *fear* versus *neutral* classification system. For each class an acoustic model is built from the data labelled as included in this class. The annotation of some segments labelled *fear* may only be due to video or contextual cues. These segments don't contain any acoustic cues of *fear*. If the proportion of such segments in the class *fear* is too high, the acoustic model of class *fear* could be too close to the model of class *neutral*.

We propose in this section to evaluate the audio cues weight to detect a situation which provokes fear. Perceptive tests carried out in a former study (Clavel et al, 2004) have already shown that emotions are perceived as more intense with the help of video support.

We thus aimed at separating segments annotated with the categorical label *fear* in the basis of audio OR video cues. Consequently, 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 a third labeller on a subcorpus. From, now on, this third labeller will be named *Lab3.* The subcorpus is composed of the segments labelled *fear* or *neutral* by at least one of the two previous labellers (see table 1). *Lab3* has to classify the segments into the categories, *fear* or *neutral.* Globally, *Lab3* annotates more segments as neutral than the two initial labellers. 54 % of the segments labelled *fear* by *Lab1,* and 43% of the segments labelled *fear* by *Lab2* are labelled *neutral* by *Lab3*. The annotation strategy of *Lab3* is closer to the annotation strategy of *Lab2* than *Lab1*.

Figure 3 and 4 show the confusion histogram between *Lab3* and *Lab1* (fig3), and between *Lab3* and *Lab2* (fig4). The annotation of *fear* class by the two initial labellers is here correlated with the intensity dimension, in order to locate the confusions with *Lab3*. As expected the confusion is higher between *fear1* (intensity level = 1) and *neutral* than between *fear3* (intensity level = 3) and *neutral.* Segments with a low level intensity are removed from the class *fear* when the "blind" annotation is considered.

The difference between the two annotation strategies emerges for the annotations of subclass *fear2* (intensity level = 2): 64% of the segments labelled *fear2* by *Lab1* are labelled *neutral* by *Lab3*. 43 % of the segments labelled *fear2* by *Lab2* are labelled *neutral* by *Lab3*. The subclass *fear2* contains fewer segments with audio cues when considering *Lab1's* annotation than when considering *Lab2's* annotation. We can thus assume that the annotation strategy of *Lab2* is more ***audio-oriented*** than the strategy of *Lab1* which is more ***context-oriented***.

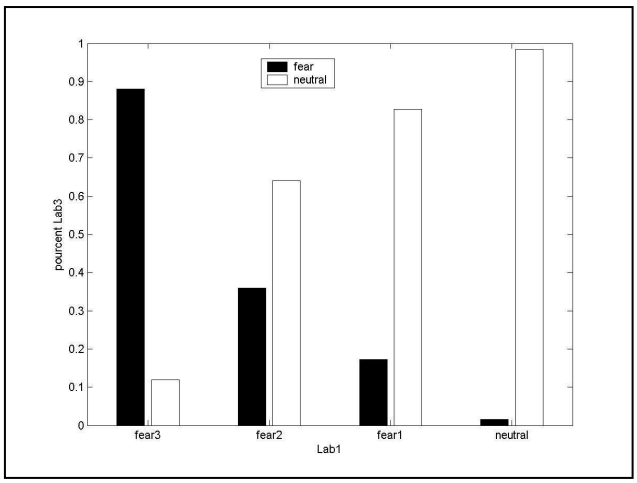

Figure 3: Confusion histogram between *Lab1* and *Lab3*

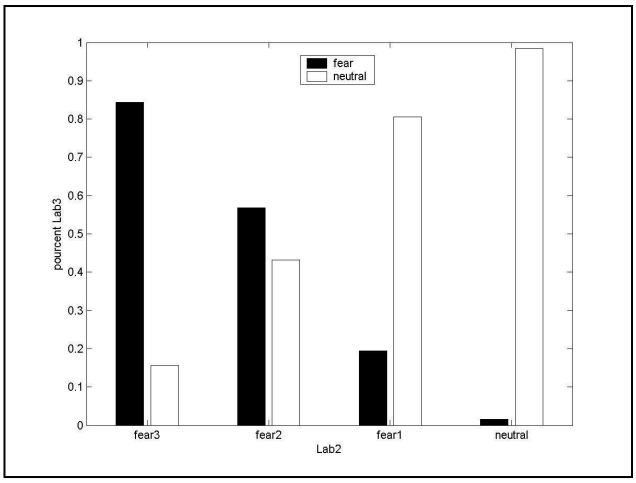

Figure 4: Confusion histogram between *Lab2* and *Lab3*

4. Perceptual versus acoustic classes

In the previous section, we compared the two initial labellers' annotation strategies. It emerges in particular that the borderline between *global fear* and *neutral* depends on the labellers strategy. In this section, we evaluate the potential influence of the various human annotation strategies on the detection system. With this purpose, we highlight the correspondence between the perceptive space delimited by each labeller and the acoustic space considered by the detection system.

4.1. Acoustic clustering

We consider here all the segments labelled *global fear* or *neutral* and evaluate the acoustic proximity between the segments without consideration of the labels. This evaluation is carried out by using the unsupervised k-means clustering (Duda & Hart, 1973). The clustering is based on the minimisation of the sum, over all clusters, of the within-cluster sums of point-to-cluster-centroïd distances. The distance used to measure the acoustic proximities between the cluster and the segment is the squared Euclidean distance. The results of this clustering are the two clusters emerging from the acoustic content regardless of the labels. We call these clusters the *acoustic clusters.*

4.2. Acoustic features vector

The emotional content is here characterized by prosodic and voice quality features (Clavel et al, 2006). The prosodic features are related to pitch (F0) and intensity contours which are extracted with Praat (Boersma & Weenink, 2005). Pitch is computed using a robust algorithm for periodicity detection based on signal autocorrelation with 40 ms frame analysis. The last prosodic feature taken into account is the duration of the voiced trajectory. The variations in terms of vocal effort are 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. 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, classically used in automatic speech recognition and used more recently for emotion detection (Shafran et al, 2003), Bark band energy and spectral centroïd (Ehrette et al, 2003) are also considered.

Features are computed every 10 ms for each segment. In order to model the temporal evolution of each feature, derivatives and statistics (min, max, range, mean, standard deviation, kurtosis, and skewness) are computed at a global level, i.e. the segment level. Each segment is then represented by a total of 174 features. All the features are normalised by their global maximum so that they are put on a single scale between -1 and 1.

The feature space is reduced by combining the different features to form 40-dimension vector (Principal Component analysis). Selection feature algorithms are not used here, as the goal is to carry out a blind clustering without any consideration on the perceptual labels of the segments.

4.3. Protocol and results

The categorical labels are correlated with the intensity dimension so that the perceptual classes considered are *neutral, fear1, fear2, fear3*.

The distribution of the segments of each label (or *perceptual class*) among the *acoustic clusters* provides an evaluation of the acoustic proximities inside *a perceptual class*. This distribution is evaluated for each annotation strategy, that is, for each labeller's annotation choice.

This analysis is performed on a subcorpus containing only *good quality* segments labelled *global fear* and *neutral*. The quality of the speech in the segments has been evaluated by the labellers. Overlaps have been avoided. Table 2 stores the number of segments used for the clustering for each labeller *Lab1* and *Lab2*. 1073 segments are used for *Lab1* and 1215 segments for *Lab2*.

	Fear 3	Fear 2	Fear 1	Neutral
Lab1	184	261	191	437
Lab2	70	226	190	729

Table 2: Repartition of the segments according to the labeller

The repartitions of the segments of each perceptual class among the two acoustic clusters, *cluster1* and *cluster2,* are considered in the two tables 3 (for *Lab1*) and 4 (for *Lab2*). This repartition is provided considering different divisions of the perceptual space, gathering for example the classes *fear2* and *fear1* or the classes *fear3* and *fear2*.

	Fear 3	Fear 2	Fear 1	Neutral
Cluster1	52%	54%	41%	46%
	50%			
Cluster2	48%	46%	59%	54%
	50%			

Table 3: *Lab1*.

	Fear 3	Fear 2	Fear 1	Neutral
Cluster1	81%	49%	40%	36%
	51%			
Cluster2	19%	51%	60%	64%
	49%			

Table 4: *Lab2*.

The acoustic borderline between the two clusters is located between *fear2* and *fear3* for *Lab2* and between *fear1* and *fear2* for *Lab1*. The trend of *Lab1* is to evaluate the segments as more intense than *Lab2*. It means for example that a same segment is evaluated *fear3* by *Lab1* and *fear2* by *Lab2*. This divergence of strategy is closely akin to the emerging position of the acoustic borderline corresponding to each labeller. In addition, the position of this borderline is more precise when considering *Lab2*'s annotation. 81% of *fear3* segments are grouped in the same acoustic cluster.

In all cases the global class *fear* emerges always less clearly in one cluster than the class *fear3*. The first cluster contains 51% of the segments labelled *fear* by *Lab2* and 64%of the segments labelled *neutral* by *Lab2*.

In order to improve the correspondence between the acoustic space and the perceptive space, the "blind" annotation is considered. We select the segments labelled *fear* or *neutral* by *Lab1*, *Lab2,* and *Lab3* at once. Table 5 shows the repartition of these labels among the two clusters. The correspondence between the acoustic clusters and the perceptual classes is here better. Now, cluster 1 contains 55% of the segments labelled *fear* and 66% of the segments labelled *neutral.*

	Fear	Neutral
Cluster1	55%	34%
Cluster2	45%	66%

Table 5: *Lab3*

5. Conclusion

In this paper we compare perceptual emotional classes to an unsupervised acoustic classification obtained by the k-means descriptive method. Perceptual classes are delimited thanks to the human annotation conducted by two initial labellers. The strategies of each labeller are first compared. It appears that the borderline between *emotion* and *neutral* depends on the annotation strategy. A third "blind" annotation highlights the role of audio information in the perception of fear-type emotions and helps at analysing the annotation strategies of the two initial labellers. Two human strategies are observed: a first one, context-oriented which mixes audio and contextual (video an temporal) information in emotion categorization; and a second one, based mainly on audio information. The k-means clustering confirms the role of audio cues in human annotation strategies. It particularly helps at evaluating those strategies from the point of view of a detection system.

First of all, the detection system is based on the detection of the global class *fear* which contains a high variability of manifestations in terms of intensity levels. The global class contains both fear-type emotions with a low level intensity such as anxiety and fear-type emotions with a high level of intensity such as terror. It emerges here that fear with a high level of intensity (*fear*3) is strongly expressed at acoustic level, successfully characterized by the acoustic features. Consequently more work needs to be done to model borderline emotions, i.e. fear1 type or some of the fear2 type.

Secondly, audio-oriented descriptions are more relevant to an emotion detection system based on audio cues. However, emotions are conveyed by other channels as well. The combination of acoustic information with other linguistic levels (lexical, dialogic, and contextual) will require context-oriented description of emotional speech. Finally, in the perspective of a multimodal surveillance system, the video cues will have to be considered and integrated in the annotation.

6. Bibliographical references

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

Clavel C., Vasilescu I, Devillers L., Ehrette T. (2004) Fiction database for emotion detection in abnormal situation, ICSLP, Jeju.

Clavel C., Vasilescu I., Richard G. and Devillers L. (2006) 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)

Clavel C., Vasilescu I., Richard G. and Devillers L. (2006) Voiced and Unvoiced content of fear-type emotions in the SAFE Corpus. Speech Prosody, Dresden – to be published.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W., & Taylor, J. (2001). Emotion recognition in human-computer interaction. IEEE Signal Processing Magazine: 18 (1), 32–80.

Craggs, R. (2004). Annotating emotion in dialogue – issues and approaches. 7th Annual CLUK Research Colloquium.

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

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.

Duda, R. O., & Hart, P. E. (1973). Pattern classification and scene analysis. New York: Wiley & Sons.

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.

Kipp, M., (2001). Anvil a generic annotation tool for mul-timodal dialogue. Eurospeech.

Lee, C.M., Narayanan, S., Pieraccini, R. (2001), Recognition of Negative Emotions from the Speech Signal. In Proceeding of the IEEE Automatic Speech Recognition and Understanding.

Osgood, C., May, W.H., Miron M. S. (1975), Cross-cultural universals of affective meaning, University of Illinois Press, Urbana.

Seigal (1988) “Non-parametric statistics”. Second Edition. Mc-Graw-Hill.

Shafran, I., Riley, M., Mohri, M., (2003). Voice Signatures. Automatic Speech Recognition and Understanding Work-shop.