PASSIVE VERSUS ACTIVE: VOCAL CLASSIFICATION SYSTEM

Z. Hammal[1], B. Bozkurt[2], L. Couvreur[2], D. Unay[2], A. Caplier[1] and T. Dutoit[2]

[1]Laboratory of images and signals LIS
46 avenue Félix Viallet, Grenoble, France
[2] Circuit Theory and Signal Processing Laboratory,
Faculté Polytechnique de Mons
1 Avenue Copernic, B-7000, Mons , Belgium

ABSTRACT

Five expressions are commonly considered to characterize human emotional states: *Happiness*, *Surprise*, *Anger*, *Sadness* and *Neutral*. Different measures can be extracted from speech signals to characterize these expressions, for example the pitch, the energy, the SPI and the speech rate. Automatic classification of the five expressions based on these features shows a great confusion between *Anger*, *Surprise* and *Happiness* on the one hand and *Neutral* and *Sadness* on the other hand. Such a confusion is also observed when humans make the same classification. We propose to define two classes of expression: *Active* gathering *Happiness*, *Surprise* and *Anger* versus *Passive* gathering *Neutral* and *Sadness*. Such a partition is also better suited for the integration of speech information in a multimodal classification system based on speech and video, which is the long term aim of our work. In this paper, we test several classification methods, namely a Bayesian classifier, a Linear Discriminant Analysis (LDA), the K Nearest Neighbours (KNN) and a Support Vector Machine with gaussian radial basis function kernel (SVM). For the considered two classes, the best performances are achieved with the SVM classifier with a recognition rate of 89.74% for *Active* state and of 86.54 % for *Passive* state.

1. INTRODUCTION

The user interface for computer systems is evolving into an intelligent multi-modal interface by taking into account the user's behaviours, speech and facial expressions in order to make the use of machines as natural as possible [1].

Vocal and facial expressions have been studied by many authors independently. The results in [2] indicate that the judgements based on a single modality depends on the type of expression to recognize and on the considered communication context.

The work presented here is in the continuity of our works on recognition of facial expressions based on video analysis [3] towards a multimodal system of expressions recognition.

Several works were carried out on recognition of vocal expressions: characteristics analysis [4,5], classification of vocal expressions [6,7]. Contrary to those works which aim at discriminating several predefined vocal expressions (from 5 to 10 different classes), our objective is to find more general and realistic vocal expressions classes. Therefore we study the statistical properties of an emotional database and try to detect broad classes of vocal expressions with consistent statistical parameter distributions.

We used the Danish Emotional Speech Database (DES) [8] which is made of the 5 currently used expressions: *Anger*, *Surprise, Happiness, Neutral* and *Sadness.* As reported in previous works [6,7], a high rate of confusion between the various classes is obtained while trying to classify them in five classes. Contrary to the common approach which try to solve these confusions by the addition of new characteristics [7], we rather consider this conflict not as a criterion of dissimilarity but as a criterion of similarity between the confused expressions. This leads to the definition of two more realistic classes: *Active* and *Passive*. Several classifiers (Bayesian classifier, LDA, KNN, SVM with gaussian radial basis function kernel) are tested to confirm the ability of the selected features to discriminate both classes.

Section 2 presents the speech database used in this work. In section 3 we present features extraction and analysis. Section 4 is dedicated to the results and discussion.

2. SPEECH DATABASE

For our experiments, we used the DES database [8]. The data were collected from two male and two female professional actors. The following expressions have been investigated : *Neutral*, *Surprise*, *Happiness*, *Sadness* and *Anger*. For each expression, there are 2 single words, 9 sentences and 2 longer passages of continuous speech. A high quality microphone was used, which did not influence the spectral amplitude or phase characteristics of the speech signal.

To check the accuracy of the simulated data, a listening test has been performed by the authors of the database to check if listeners (20 normal-hearing, 10 of each gender) could identify the emotional content of the recorded utterances. The utterances were correctly identified with an average rate of 67% (cf.Table1). *Surprise* and *Happiness* were often confused as well as *Neutral* and *Sadness*.

	Neutral	Surprise	Happiness	Sadness	Anger
Neutral	**60.8**	2.6	0.1	31.7	4.8
Surprise	10.0	**59.1**	28.7	1.0	1.3
Happiness	8.3	29.8	**56.4**	1.7	3.8
Sadness	12.6	1.8	0.1	**85.2**	0.3
Anger	10.2	8.5	4.5	1.7	**75.1**

Table.1 Confusion matrix from subjective human evaluation [8]. Columns represent the vocal expression selected for utterances for the vocal expressions input of each row.

3. FEATURE EXTRACTION AND ANALYSIS

Based on recent studies of vocal expressions analysis, several prosodic features have been defined [4,5,9]. Guided by the works [4,5,6,7] we restrict ourselves to the following features: the pitch, the energy, the SPI and the speech rate.
Analysis is carried out to extract pitch, energy and SPI for constant length (30msec) and constant shift (10msec) speech frames extracted from recordings. Next we performed a statistical analysis in order to select the acoustical parameters that could display the differences between vocal expression categories.
Our analysis was guided by the works [4,5,6] on the correlation between the sets of characteristics extracted for each parameter and the vocal expressions.
In the following we are interested only in the normalized characteristics (zero-mean and standard deviation to 1) presented in Table.2.

	Range	Median	Standard deviation	Rises	Falls	Max
F0	x	x	x	x	x	x
Energy	x	x	x	x	x	x
SPI	-	-	-	-	-	x
Speech rate						

Table.2 Statistical parameters used for each characteristic. ‘x’: used, ‘-‘: not used.

3.1 The Pitch

The pitch (F0) is the fundamental frequency of the acoustic signal. This feature is computed using an autocorrelation based pitch estimator [10]. Statistics related to F0 such as minimum, maximum, mean, median, range, standard deviation are computed. Flatness of intonation is also measured with two values: median values of the rises and falls of F0 [9].
Fig.1 presents the result of the range (maximum-minimum), median and standard deviation for F0. The value of each bar corresponds to the mean value for all the data for each expression. The standard deviation of this value is also reported. Fig.2 presents the median of the rises and of the falls of F0 for every expression.
Fig.1 and Fig.2 show that two groups of expressions appear: the statistical values of F0 for *Surprise*, *Anger* and *Happiness* examples are comparatively higher than the corresponding values for *Neutral* and *Sadness* examples.

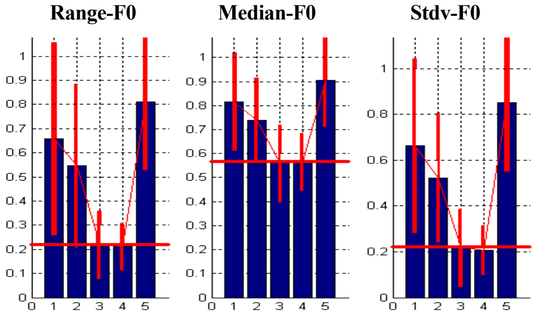

Fig.1 Mean values of range, median and standard deviation of F0 for all the data and all the expressions. The bars represent the expressions in the following order : 1) *Anger*, 2) *Happiness*, 3) *Neutral*, 4) *Sadness*, 5) *Surprise*.

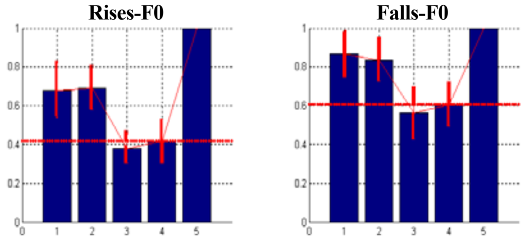

Fig.2 Mean values of rises and falls for F0 for all the data and all the expressions. The bars represent the expressions in the following order : 1) *Anger*, 2) *Happiness*, 3) *Neutral*, 4) *Sadness*, 5) *Surprise*.

3.2 The Energy

The signal energy is computed (in decibels) as sum of square of the discrete signal [10]. We compute the energy of only the voiced segment in utterances to avoid jumps at plosives. Similarly as applied to the pitch, we compute a set of global statistics such as minimum, maximum, median, range, standard deviation and medians of rises and falls.
Fig.3 and Fig.4 present the statistical characteristics of energy. *Sadness* and *Neutral* speech show lower range, median, standard deviation, rises and falls compared to the other vocal expressions. These results are coherent with the fact that *Anger*, *Happiness* and *Surprise* require more energy than *Neutral* and *Sadness* expressions [9].

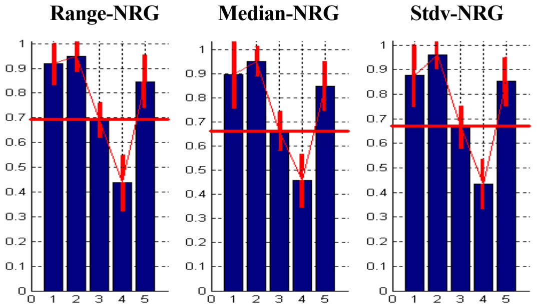

Fig.3 Mean values of range, median and standard deviation of energy for all the data and all the expressions. The bars represent the expressions in the following order : 1) *Anger*, 2) *Happiness*, 3) *Neutral*, 4) *Sadness*, 5) *Surprise*.

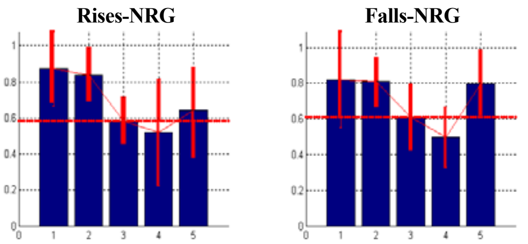

Fig.4 Mean values of rises and falls for energy for all the data and all the expressions. The bars represent the expressions in the following order : 1) *Anger*, 2) *Happiness*, 3) *Neutral*, 4) *Sadness*, 5) *Surprise*.

3.3 The SPI

SPI is a spectral measure of the ratio of low-frequencies (70-1600HZ) to the high-frequencies (1600-4500Hz) [11]. It is used as a simple approximation of the ``harshness'' vs. ``softness'' of the voice quality in the area of speech therapy.

The analysis of the characteristics of the SPI of voiced frames shows that only the maximum value is relevant for classification. This value presents the same behaviour as F0 and the energy (Fig.5).

3.4 The speech rate

Speech rate is computed for each recording as the number of phonemes spoken in a given time interval. Since the text content is known, the number of phonemes of each recording is available in the database. Only the estimation of speech duration from the recorded signal is necessary for this task. Recordings are segmented into speech and non-speech segments by applying an energy threshold for decision. The energy threshold is defined proportionally to the mean energy of each recording.

The analysis of the speech rate (Fig.5) shows that this feature is higher for *Surprise*, *Anger* and *Happiness* than for *Neutral* and *Sadness*.

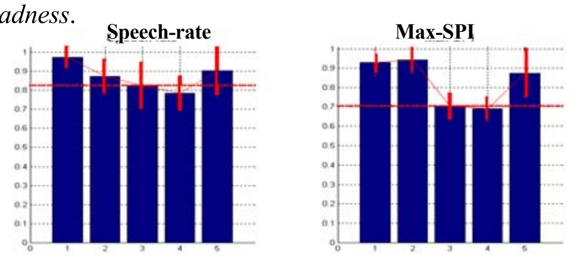

Fig.5 Speech rate mean values and SPI maximum mean values for all the data and all the expressions. The bars represent the expressions in the following order : 1) *Anger*, 2) *Happiness*, 3) *Neutral*, 4) *Sadness*, 5) *Surprise*.

3.5 Conclusion of the analysis

Expressions such as *Anger*, *Surprise*, *Happiness* have higher values of F0, energy, SPI and speech rate which means that they are related to a strong activity. On the contrary, expressions such as *Sadness* or *Neutral* present smaller values of F0, energy and SPI as well as a decrease of speech rate, which means that they are related to a small activity. These observations lead us to conclude that the frequently used acoustic parameters do not discriminate the 5 considered vocal expressions. Moreover these parameters exhibit two groups of expressions: *Anger*, *Surprise* and *Happiness*, on the one hand, and *Sadness* and *Neutral*, on the other hand.

4. RESULTS AND DISCUSSION

4.1 Five expressions classification

The analysis of section 3 shows that there are prosodic similarities between several vocal expressions. We first want to confirm these results and to see the effect of these characteristics on the discrimination between the vocal expressions. We use the 12 global acoustic features described in section 3 (median, range, standard deviation, rises and falls for F0 and energy, maximum for SPI and speech rate) with a Bayesian classifier to classify the five vocal expressions. In order to minimize the effect of the lack of data we use a bootstrap method [12] to better estimate the classification rates. It consists in duplicating the number of data by random pulling with handing-over. In our case 50 databases are built from the initial database. Classification rates are calculated in the following way: at each iteration, we train the classifier on one database and we test on the initial database. The process is reiterated on the 50 databases. The final rates are the mean of the 50 rates.

Table 3 presents the results of our classification. To test the validity of our characteristics we compare these results with those of [7]. By using more characteristics, the classification rate obtained in [7] is around 50% (*Neutral* (51%), *Surprise* (64%), *Happiness* (36%), *Sadness* (70%) *and Anger* (31%)) whereas ours is around 54%. A more important observation is that our results are more homogeneous: the classification rate is almost the same for all the expressions which is not the case in [7].

	Neutral	Surprise	Happiness	Sadness	Anger
Neutral	**46.76**	23.92	12.26	3.3	13.73
Surprise	20.11	**51.69**	6.5	5	16.61
Happiness	7.11	5	**56.61**	24.69	6.5
Sadness	4.57	3.19	28.76	**61.80**	1.65
Anger	12.5	29.11	4.26	1.84	**52.26**

Table.3 Confusion matrix with a Bayes classifier.

4.2 *Passive* versus *Active* classification

Comparison with Table 1 indicates that human listeners show the same tendency. The emotions that have been confused are those with similar acoustic characteristics (see section 3). Considering the confusions as indicators of the similarity perceived between the confused expressions we decide to create two classes: *Active* which includes *Anger*, *Happiness* and *Surprise* and *Passive* which includes *Neutral* and *Sadness*.

To be sure of the discrimination between the two new classes, we compare the classification rates obtained with 4 different classifiers: the Bayesian classifier, the Linear Discriminant Analysis (LDA) [13], the K nearest neighbours (KNN) with 5 neighbours and Euclidian metric [13] and the Support Vector Machine with gaussian radial basis function kernel (SVM) [14].

The classification rates are obtained by 5-fold cross validation. The results of classification (Tables 4-7) show that the recognition rates of Bayesian classifier and LDA are lower than SVM and KNN. This is due to the fact that Bayesian classifier assumes Bayesian distributions of classes, which may be a false assumption for our dataset, and LDA performs a linear separation while our data may be non-linear.

The KNN performs better result than LDA, however SVM gives the best classification rates (Table.7).

The presented results (Table 6-7) makes it possible to confirm that the characteristics used are sufficient for our two classes classification.

	Active	Passive
Active	**78.84%**	21.15%
Passive	19.23%	**80.76%**

Table.4 Results of Bayesian classification.

	Active	Passive
Active	**96.79%**	3.2%
Passive	46.15%	**53.85%**

Table.5 Results of LDA classification.

	Active	Passive
Active	**83.33%**	16.67%
Passive	11.54%	**88.46%**

Table.6 Results of KNN classification.

	Active	Passive
Active	**89.74%**	10.26%
Passive	13.46%	**86.54%**

Table.7 Results of SVM classification.

5. CONCLUSION

In order to integrate speech modality to expressions classification system based on video, we investigated acoustic properties of speech associated with five different vocal expressions (*Surprise*, *Happiness*, *Anger*, *Neutral* and *Sadness*). The analysis of the acoustics features enables to note that the considered acoustic features provide rather limited support to separate the five vocal expressions. However results show that grouping expressions into two larger classes according to the statistical parameters derived from acoustic features results in successful classification: *Happiness*, *Anger* and *Surprise* in one class and *Neutral* and *Sadness* in the other. The same confusions are found for a classification by humans, which leads us to define two classes of vocal expressions: *Active* and *Passive*. Classification rates are very satisfactory. The interest of this classification is that it is more compliant with real applications.

The development of a multimodal expressions recognition system is under study. Such a system will combine at the same time both modalities (video and speech) for better recognition or will use them separately according to the context of the application.

Acknowledgement :

This work was supported by the Similar network of excellence.

REFERENCES

[1] http://www.similar.cc

[2] P. Ekman, W.V. Friesen, M. O'Sullivan & K. Scherer. “Relative importance of face, body, and speech in judgments of personality and affect”. Journal of Personality and Social, Psychology, vol 38(2), pp. 270-277, 1980.

[3] Z. Hammal, L. Couvreur, A. Caplier, M. Rombaut, “Facial Expressions Recognition Based on The Belief Theory: Comparison with Different Classifiers”, in Proc. 13-th International Conference on Image Analysis and Processing, ICIAP, 2005.

[4] K. R. Scherer, “Vocal communication of emotion” A review of research paradigms. Speech Communication, 2003, vol 40, pp. 227-256.

[5] P. N. Juslin, P. Laukka “Communication of emotions in vocal expression and music performance: Different channels, same code?” Psychological Bulletin, 2003, pp. 770-814.

[6] V. A. Petrushin “Emotion recognition in speech signal: experimental study, development, and application”, in Proc. 6-th International Conference on Spoken Language processing, ICSLP, 2000.

[7] D. Ververidis, C. Kotropolos “Automatic speech classification to five emotional states based on gender information”, in Proc. Eusipco, Vienna , 2004, pp.341-344.

[8] I. S. Engberg, A. V. Hansen, “Documentation on the Danish Emotional Speech Database DES”, Alborg, September, 1996.

[9] M. Schröder, “Speech and Emotion Research”, PHD thesis report, university of Saarlandes, 2003.

[10] T. F. Quatieri, “Discrete Time Speech Signal Processing: Principles and Practice”, Prentice Hall PTR, 2001.

[11] D. Deliyski,, “Acoustic model and evaluation of pathological voice production”, in Proc. 3-rd Conference on Speech Communication and Technology EUROSPEECH , Berlin, Germany, 1993, pp.1969-1972.

[12] R. Kallel, M. Cottrell, V. Vigneron “Bootstrap for neural model” , Neurocomputing, 2002, vol 48, pp.175-183.

[13] R. O. Duda, P. E. Hart and D. G. Stork, “Pattern Classification” (2nd ed.), New York, John Wiley and Sons, 2001.

[14] J.C . Burges “A Tutorial on Support Vector Machines for Pattern Recognition”, Data Mining and Knowledge Discovery, Kluwer Academic Publishers, Boston. Manufactured in The Netherlands, 1998, pp. 121–167.