sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Hierarchical Classifiers Approach for Emotions Recognition"
XIII Reunión de Trabajo en Procesamiento de la Información y Control (RPIC 2009), sep, 2009.

Hierarchical Classifiers Approach for Emotions Recognition

Enrique M. Albornoz[1,2] Diego H. Milone[1,2] Hugo L. Rufiner[1,2,3]

[1] *Centro de I+D en Señales, Sistemas e INteligencia Computacional (SINC(i))*
Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral
[2] *Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET)*
[3] *Laboratorio de Cibernética, Facultad de Ingeniería, Universidad Nacional de Entre Ríos*

***Abstract*— The recognition of the emotional states of speaker is a multi-disciplinary research area that has received great interest in the last years. One of the more important goals is to improve the voiced-based human-machine interactions. Recent works on this domain use the prosodic features and the spectrum characteristics of speech signal, with standard classifier methods. However, there is no analysis of what are the causes of the results obtained. Furthermore, for traditional methods the improvement in performance has also found a limit. In this paper, a study of spectral characteristics of emotional signals is presented. This information is also used in order to group emotions based on their spectral similarities. Hidden Markov Models and Multilayer Perceptron have been evaluated in different configurations with different features, to design a new hierarchical method for emotions classification. Results with the hierarchical method improve up to 6.35% the recognition rate.**

***Keywords*— Emotion Recognition, Spectral information, Hierarchical Classifiers, Hidden Markov Model, Multilayer Perceptron.**

1. INTRODUCTION

In the last years, the recognition of emotions has become in a multi-disciplinary research area that has received great interest. This plays an important roll in the improvement of human-machine interaction. For example, security application of the fear emotional manifestation in abnormal situations is studied in [5]; in [8], real-life emotion detection using a corpus of real agent-client spoken dialogs from a medical emergency call center is studied; in [22], a framework to support semi-automatic diagnosis of psychiatric diseases is proposed.

The use of biosignals (like ECG, EEG, etc.), face images and body images is an interesting alternative to detect emotional states [19, 12, 23]. However, methods to record and use these signals are more invasive, complex and not possible in some real applications. Therefore, the use of speech signals clearly becomes a feasible option. Most of the previous works in emotion recognition have been based in the analysis of speech prosodic features and spectral information. For the classifier, Hidden Markov Models (HMM) and several other standard techniques have been explored for this task [3, 9, 17].

Very few works have been presented using some combination of traditional standard methods. In [14], two classification methods: stacked generalization and unweighted vote, were applied in emotion recognition. This classifier improved the performance of traditional classification methods. In [21], a multiple stages classifier with Support Vector Machines (SVM) is presented. Two class decision is repetitively made until only one class remains, and hardly separable classes are divided at last. Authors build this partition based on expert knowledge or derived it from the confusion matrices of a SVM approach. A two stages classifier for five emotions is proposed in [13]. In this work, a SVM to classify five emotions into two groups is used. Then, HMMs are used to classify emotions within each group.

In this work, an analysis of spectral features is made in order to define groups of similar emotions. Emotions are grouped based on their properties and a hierarchical classifier is designed. The proposed classifier is evaluated in the same experimental condition than standard classifiers, with important improvements in the recognition rates.

In the next section, the emotional speech data base and an acoustical analysis of emotions are presented. Section 3 describes the feature extraction and classification methods. The method here proposed and the experiments are also explained. Section 4 deals with the classification performance and discussion. Finally, conclusions and future works are presented.

2. ACOUSTIC ANALYSIS OF EMOTIONS

2. 1. Emotional Speech Corpus

The emotional speech signals used were taken from an emotional speech data base [4], developed by the Communication Science Institute of Berlin Technical University. This corpus had been used in several studies [3, 9, 20] and allows the development and evaluation

Table 1: Distribution of emotions in the corpus.

Emotion	Number of utterances
Anger	127
Boredom	81
Disgust	46
Fear	69
Joy	71
Sadness	62
Neutral	79

of an speaker independent recognizer[1].

The corpus, consisting of 535 utterances, includes sentences performed in 6 ordinary emotions, and sentences in neutral emotional state. These emotions covers the "big six" emotions set except for boredom instead of surprise. Sentences are labeled as: happiness (joy), anger, fear, boredom, sadness, disgust and neutral (Table 1 shows their distribution).

The same texts were recorded in german by ten professional actors, 5 female and 5 male, which allows studies over the whole group, comparisons between emotions and comparisons between speakers. The corpus consists of 10 utterances for each emotion type, 5 short and 5 longer sentences, from 1 to 7 seconds. To achieve a high audio quality, these sentences were recorded in an anechoic chamber with 48 kHz sample frequency (later downsampled to 16 kHz) and were quantized with 16 bits per sample. A perception test with 20 subjects was carried out to ensure the emotional quality and naturalness of the utterances.

2.2. Acoustic Analysis

The psychological conceptualization of affects, with two-dimensional and three-dimensional models, is widely known in the categorization of emotions [6, 11, 18]. These models are often used to group emotions in order to define classes. For example, those associated with low arousal and low pleasure versus those associated with high arousal and high pleasure, etc. In this work the psychological information will be discarded and emotions would be characterized by spectral information.

For every utterance, the mean of the log-spectrum (MLS) on each frequency band, frame by frame, were calculated. Then, the average of the mean log-spectrums (AMLS) over all the emotions with same class, for every classes, were computed

$$AMLS_k(f) = \frac{1}{N_k} \sum_{i=1}^{N_k} \frac{1}{T_i} \sum_{t=1}^{T_i} \log \|v(t,f)\| \qquad (1)$$

where f is a frequency band, N_k as the number of sentences for the emotion class k, T_i is the number of temporal windows in the utterance i and $v(t,f)$ is the discrete Fourier transform of the signal in the frame t.

The most important information to discriminate between emotion classes was found between 0 and 1200 Hz. In Figure 1, this information is shown for each emotional class. As it can be seen in the figures, some emotions have spectral similarities between them. For example, it can be noticed a similar shape and a maximum between 240 and 280 Hz in Joy, Anger and Fear. A minimum is present close to 75 Hz in Joy, Anger, Fear and Disgust. On the other hand, Boredom, Neutral and Sadness have similar shape and a peak between 115 and 160 Hz.

So, it is possible to define groups using the spectral information. For example, a group could contain Joy, Anger, Fear emotions whereas other contains Boredom, Neutral and Sadness emotions and finally Disgust emotion alone in a third group. On the other hand, emotion similarities used to propose the groups keep a relationship with accuracies and errors present in confusion matrices [3, 9, 15]. This relevant knowledge for emotion grouping will be used in the next section to define a hierarchical classifier.

3. PROPOSED METHOD

In this section, a new hierarchical classification method based on the acoustic analysis described above is presented. In order not to favor one of the emotions over the others, in the experiments the same number of utterances was used for every emotion. This balanced partition has 46 randomly selected utterances for each emotion. Every utterance has one label according to the expressed emotion and represents only one pattern.

3.1. Features Extraction and Classification Methods

For every emotion utterance, three kinds of characteristics were extracted: MLS, mel frequency cepstral coefficients (MFCCs) and prosodic features. The spectrograms were calculated by using Hamming windows of 25 ms with a 10 ms frame shift. The MLS were computed as defined in Section 2.2. The same windowing parameters were used to obtain the MFCC parametrization. The first 12 MFCC plus the first and second derivatives were extracted [24].

The use of prosodic features has been discussed above and classic methods to calculate the *Energy* and the F_0 along the signals have been used [7]. Many parameters can be extracted from them; therefore the minimum, mean, maximum and standard deviation, over the whole utterances, were used. This set of parameters has been already studied and the experiments reported an important information gain to distinguish emotions [1, 3, 21]. Combinations of features were arranged in vectors and every dimension was independently normalized by the maximum from the feature vectors set.

To take advantage of spectral similarities revealed in Section 2.2 analyses, here an alternative method to the standard classifier (Fig. 2) is presented. Classifiers are

[1] The information is accessible from http://pascal.kgw.tu-berlin.de/emodb/.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Hierarchical Classifiers Approach for Emotions Recognition"
XIII Reunión de Trabajo en Procesamiento de la Información y Control (RPIC 2009), sep, 2009.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Hierarchical Classifiers Approach for Emotions Recognition"
XIII Reunión de Trabajo en Procesamiento de la Información y Control (RPIC 2009), sep, 2009.

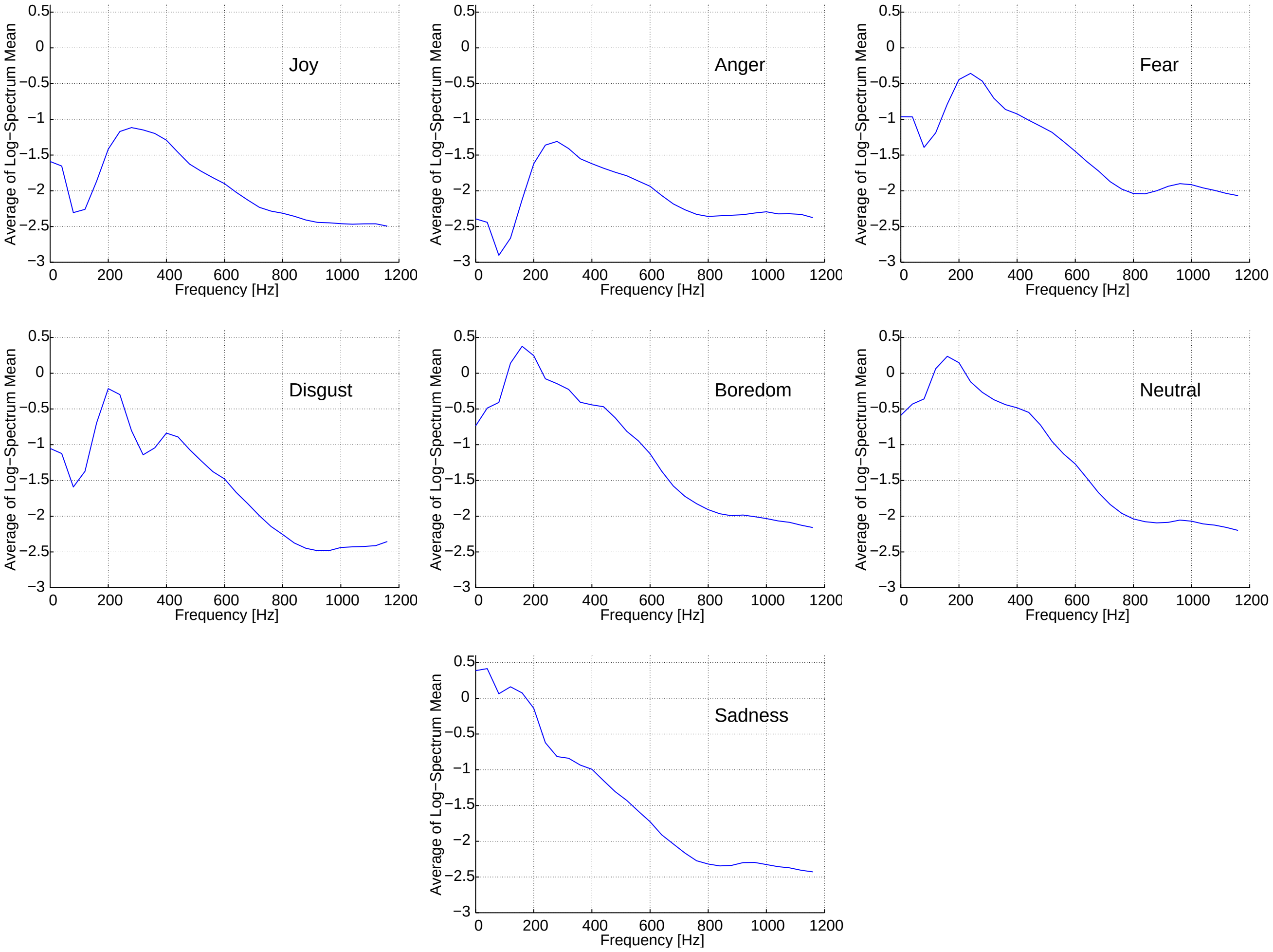

Figure 1: Average of Mean Log-Spectrum for all emotion classes.

based on two well known techniques: Multilayer Perceptron (MLP) and Hidden Markov models (HMM). The MLP is a class of Artificial Neural Network and it consists of a set of process units (simple perceptrons) arranged in layers. Often the nodes are fully connected between layers without connections between units in the same layer. The input vector (feature vector) feeds into each of the first layer perceptrons, the outputs of this layer feed into each of the second layer perceptrons, and so on [10].

The HMMs are basically statistical models that describe sequence of events and it is a very used technique in speech and emotions recognition. In classification tasks, a model is estimated for every signal type. Thus, it would take into account as many models as signal classes to recognize. During classification, the probability for each signal given the model is calculated. The classifier output is based on the model with the maximum probability of generating the signal [16].

Based in previous studies [2], a two state HMM was defined because it achieved the best results. Tests increasing the number of Gaussian components in the mixtures by two every time, were performed to find the optimal structure. In order to optimize the MLP performance, different number of neurons in hidden layer were tested.

3. 2. Hierarchical Classifiers

The main motivations for the development of a hierarchical classifier are to take advantage of spectral emotion similarities, to improve the emotion recognition rate and to use the fact that better results can be reached when the number of emotions decrease for the same standard classifier. As can be seen from Fig. 3,

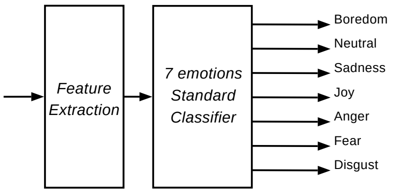

Figure 2: Standard classifier for 7 emotions.

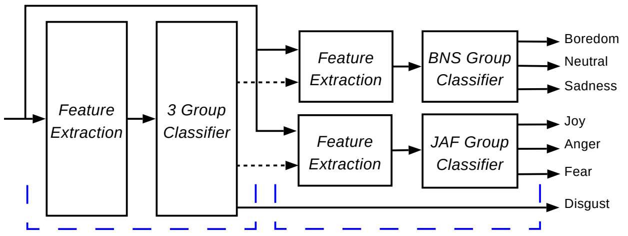

Figure 3: General structure of the hierarchical classifier for 7 emotions.

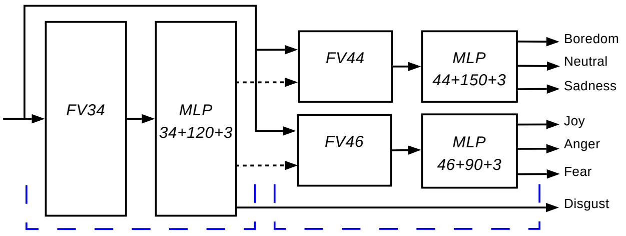

Figure 4: Best hierarchical classifier for 7 emotions.

the hierarchical classifier is defined in two stages. In a first stage the emotion utterance would be classified in one of 3 groups (BNS, JAF or Disgust), then it would be classified again in the corresponding block if it is not Disgust and finally the emotion label is obtained.

To define the hierarchical model structure in each block, diverse configurations of MLP and HMM with different parameter vectors, were evaluated. Finally, the model stages were chosen and assembled with classifiers that achieved better results in isolated block tests.

In every MLP block test, 15 feature vectors were used in 3 different hidden layer configurations (90, 120 and 150 perceptrons). Table 2 shows the number of characteristics for each vector and what kind of features it includes. For example, the feature vector FV14 includes 12 MFCC, the F_0 mean and the Energy mean. On the other hand, a 36 coefficients vector was used for HMM tests (12 MFCCs plus delta and acceleration), like in [2].

In MLP experiments, 60% of data was randomly selected for training, 20% was used for the generalization test and the remaining 20% was left for validation. The MLP training was stopped when the network reached the generalization point with test data [10]. In HMM cases, the 20% used for test was added to the standard train set.

4. RESULTS AND DISCUSSIONS

A comparative analysis between Gaussian Mixture Models and HMM for recognition of seven emotions was presented in [2]. Better results were achieved with a two state HMM with mixtures of 30 Gaussians, using a MFCC parametrization including delta and acceleration coefficients. Here, the same system with the balanced partitions was tested in order to obtain a baseline to compare results. The classification rate was 66.67%. In this work, the number of output nodes in the MLP equals the number of seven emotions and the performance was 68.25% for the network composed by 46 input neurons and 90 hidden neurons (considered here as the baseline for MLP classification).

For the Stage I, three different options were evaluated: (a) to group HMM baseline outputs into 3 groups (HMM^7); (b) to model each group with one HMM (HMM^3); and (c) to use MLP with 3 output neurons. In HMM cases, the number of Gaussian components in the mixture was set in 30 (as in [2]). Table 3 shows the MLP results for each feature vector with train and validation data.

Best results obtained for Stage I are summarized in Table 4. It can be seen that MLP achieved the best result but it is the worst classifying Disgust. This could be because MLP is not a good classifier when the classes are unbalanced.

For each block in Stage II, HMM and MLP tests were done using all partition data to evaluate the blocks in an isolated form. In HMM case, tests altering the number of Gaussian components in the mixture, increasing by two every time, were performed. A HMM with 26 Gaussians in the mixtures achieved a 74.07% for JAF test, while only 4 Gaussians achieved a 77.78% for the BNS case. The MLP results for JAF and BNS classification could be seen in Table 5 and Table 6 respectively. Best results for the isolated blocks of Stage II are shown in Table 7.

No blocks were rejected and the 12 combinations with the best blocks were evaluated. In Table 8, the performance for JAF and BNS blocks with both models are shown for each model in Stage I. The per-

Parameters	FV12	FV14	FV16	FV18	FV20	FV30	FV32	FV34	FV36	FV38	FV42	FV44	FV46	FV48	FV50
12 MFCC	●	●	●	●	●						●	●	●	●	●
30 Mean Log-Spectrum						●	●	●	●	●	●	●	●	●	●
$\mu(F_0)$, $\mu(E)$		●	●	●	●		●	●	●	●		●	●	●	●
$\sigma(F_0)$, $\sigma(E)$			●		●			●		●			●		●
$Min(F_0)$, $Max(F_0)$				●	●				●	●				●	●
$Min(E)$, $Max(E)$				●	●				●	●				●	●

Table 2: Feature vectors used in MLP tests.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Hierarchical Classifiers Approach for Emotions Recognition"
XIII Reunión de Trabajo en Procesamiento de la Información y Control (RPIC 2009), sep, 2009.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Hierarchical Classifiers Approach for Emotions Recognition"
XIII Reunión de Trabajo en Procesamiento de la Información y Control (RPIC 2009), sep, 2009.

Table 3: Results of MLP classification for 3 Groups. Classification rate in [%].

Input	Best Net	Train	Validation
FV12	12+90+3	98.98	85.71
FV14	14+90+3	95.92	87.30
FV16	16+90+3	97.96	87.30
FV18	18+150+3	98.47	79.37
FV20	20+90+3	100.00	77.78
FV30	30+90+3	100.00	87.30
FV32	32+90+3	99.49	85.71
FV34	34+120+3	**98.98**	**88.89**
FV36	36+90+3	99.49	84.13
FV38	38+120+3	100.00	82.54
FV42	42+120+3	92.86	87.30
FV44	44+150+3	96.94	84.13
FV46	46+150+3	94.39	85.71
FV48	48+90+3	100.00	80.95
FV50	50+150+3	100.00	82.54

Table 4: Performance of Stage I classification models.

	HMM grouped	HMM	MLP
JAF	88.89	77.78	88.89
BNS	85.19	92.59	100.00
D	66.67	88.89	55.56
average	84.13	85.71	**88.89**

Table 5: Results of JAF with MLP in isolated classification. Classification rate in [%].

Input	Best Net	Train	Validation
FV12	12+150+3	98.81	81.48
FV14	14+150+3	90.48	85.19
FV16	16+150+3	95.24	74.07
FV18	18+90+3	86.90	74.07
FV20	20+120+3	85.71	77.78
FV30	30+90+3	98.81	77.78
FV32	32+120+3	100.00	70.37
FV34	34+90+3	100.00	77.78
FV36	36+120+3	76.19	74.07
FV38	38+90+3	73.81	77.78
FV42	42+90+3	100.00	81.48
FV44	44+90+3	100.00	85.19
FV46	46+90+3	**100.00**	**85.19**
FV48	48+90+3	100.00	85.19
FV50	50+150+3	100.00	85.19

Table 6: Results of BNS with MLP in isolated classification. Classification rate in [%].

Input	Best Net	Train	Validation
FV12	12+90+3	84.52	66.67
FV14	14+90+3	100.00	74.07
FV16	16+90+3	100.00	66.67
FV18	18+150+3	96.43	48.15
FV20	20+150+3	94.05	51.85
FV30	30+90+3	100.00	74.07
FV32	32+120+3	92.86	74.07
FV34	34+90+3	96.43	66.67
FV36	36+90+3	92.86	62.96
FV38	38+120+3	100.00	59.26
FV42	42+150+3	96.43	70.37
FV44	44+150+3	**97.62**	**81.48**
FV46	46+120+3	100.00	77.78
FV48	48+90+3	95.24	59.26
FV50	50+120+3	97.62	66.67

Table 7: Best results for isolated Stage II classification.

Group	Stage II model	Performance
JAF	MLP	85.19
	HMM	74.07
BNS	MLP	81.48
	HMM	77.78

Table 8: Final test of hierarchical model.

Stage I		Stage II				
Model	Disgust	JAF		BNS		Best
		HMM	MLP	HMM	MLP	
HMM^7	66.67	66.67	**74.07**	62.96	70.37	71.43
HMM^3	**88.89**	55.56	62.96	74.07	77.78	73.02
MLP	55.56	66.67	**74.07**	77.78	**81.48**	**74.60**

formance for the best combination considering each model in Stage I are: 71.43% for HMMs grouped (HMM^7), 73.02% for 3 HMMs (HMM^3) and 74.60% for MLP.

Finally, the best hierarchical model is formed by a MLP with FV34 and 120 hidden neurons in the Stage I. A MLP with FV46 and 90 hidden neurons for the JAF block and a MLP with FV44 and 150 hidden neurons for the BNS block. Figure 4 shows the best hierarchical model configuration. As can be seen, spectrum and some prosodic features (FV34) are the best to classify the 3 groups. However, the MFCC are required to improve the recognition in both blocks of the Stage II.

5. CONCLUSIONS AND FUTURE WORKS

In this paper a characterization of emotions and their similarities based on the acoustical features was presented. A new hierarchical method for emotion classification supported by such acoustic analysis was proposed. Tests with different number of inputs and internal structure for MLP, models and tests increasing the number of Gaussians in mixtures for HMMs were performed for each block. The preliminary results show that the hierarchical model exceeds the standard (non-hierarchical) classifiers. The prosody combined with spectral features improves the results in the emotion recognition task.

In a future works will be important to do a cross-validation test with more data for the hierarchical model. Although the speaker independent results are good, tests in gender dependent frameworks are planned. This will allow taking more advantage of specific spectral information. Also, it is planned to carry out similar analyses on other languages.

6. ACKNOWLEDGEMENTS

The authors wish to thank: the *Agencia Nacional de Promoción Científica y Tecnológica* and *Universidad Nacional de Litoral* (with PAE 37122, PAE-PICT 00052, CAID 012-72), the *Universidad Nacional de Entre Ríos* (with PID 61111-2 and 6107-2), the *Consejo Nacional de Investigaciones Científicas y Técnicas* (CONICET) from Argentina, for their support.

References

[1] J. Adell Mercado, A. Bonafonte Cávez, and D. Escudero Mancebo. Analysis of prosodic features: towards modelling of emotional and pragmatic attributes of speech. In *Proc. del lenguaje natural*, number 35, pages 277–283, sep. 2005.

[2] E. M. Albornoz, M. B. Crolla, and D. H. Milone. Recognition of emotions in speech. In *Proc. of XXXIV CLEI*, Santa Fe (Argentina), sep. 2008.

[3] M. Borchert and A. Dusterhoft. Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. *NLP-KE '05. Proc. of IEEE Int. Conf. on*, pages 147–151, Oct. 2005.

[4] F. Burkhardt et al. A Database of German Emotional Speech. *Proc. Interspeech 2005*, pages 1517–1520, September 2005.

[5] C. Clavel et al. Fear-type emotion recognition for future audio-based surveillance systems. *Speech Commun.*, 50(6):487–503, 2008.

[6] R. Cowie and R. Cornelius. Describing the emotional states that are expressed in speech. *Speech Comm.*, 40(1):5–32, 2003.

[7] J. R. Deller, J. G. Proakis, and J. H. Hansen. *Discrete-Time Processing of Speech Signals*. Macmillan Publishing, New York, 1993.

[8] L. Devillers and L. Vidrascu. *Speaker Classification II*, volume 4441/2007 of *LNCS*, chapter Real-Life Emotion Recognition in Speech, pages 34–42. Springer, Berlin, 2007.

[9] M.M.H. El Ayadi, M.S. Kamel, and F. Karray. Speech Emotion Recognition using Gaussian Mixture Vector Autoregressive Models. *ICASSP 2007. IEEE Int. Conf. on*, 4:957–960, April 2007.

[10] S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 2 edition, July 1998.

[11] J. Kim. *Robust Speech Recognition and Understanding*, chapter Bimodal Emotion Recognition using Speech and Physiological Changes, pages pp. 265–280. I-Tech Education and Publishing, Vienna, Austria, 2007.

[12] J. Kim and E. André. Emotion recognition based on physiological changes in music listening. *Pattern Analysis and Machine Intelligence, IEEE Tran. on*, 30(12):2067–2083, Dec. 2008.

[13] Yi-Lin Lin and Gang Wei. Speech emotion recognition based on HMM and SVM. *Machine Learning and Cybernetics, 2005. Proc. of Int. Conf. on*, 8:4898–4901, aug. 2005.

[14] D. Morrison, R. Wang, and L. C. De Silva. Ensemble methods for spoken emotion recognition in call-centres. *Speech Comm.*, 49(2):98 – 112, 2007.

[15] A. Noguerias et al. Speech Emotion Recognition Using Hidden Markov Models. *Eurospeech 2001*, pages 2679–2682, 2001.

[16] L. R. Rabiner and B. H. Juang. *Fundamentals of Speech Recognition*. Prentice-Hall, 1993.

[17] J. Rong et al. Acoustic Features Extraction for Emotion Recognition. *ICIS 2007. 6th IEEE/ACIS Int. Conf. on*, pages 419–424, July 2007.

[18] K. R. Scherer. What are emotions? And how can they be measured? *Social Science Information*, 44(4):695–729, Dec. 2005.

[19] K. Schindler, L. Van Gool, and B. de Gelder. Recognizing emotions expressed by body pose: A biologically inspired neural model. *Neural Networks*, 21(9):1238–1246, 2008.

[20] B. Schuller et al. Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition. *Multimedia and Expo, 2008 IEEE Int. Conf. on*, pages 1333–1336, April 2008.

[21] B. Schuller, G. Rigoll, and M. Lang. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. *Proc. IEEE-ICASSP '04*, pages I–577–80 vol.1, May 2004.

[22] D. Tacconi et al. Activity and emotion recognition to support early diagnosis of psychiatric diseases. *PervasiveHealth 2008*, pages 100–102, Feb. 2008.

[23] V. Vinhas, L. P. Reis, and E. Oliveira. Dynamic Multimedia Content Delivery Based on Real-Time User Emotions. Multichannel Online Biosignals Towards Adaptative GUI and Content Delivery. In *BIOSIGNALS 2009*, pages 299–304, Porto (Portugal), 2009.

[24] S. Young et al. *The HTK Book (for HTK Version 3.1)*. Cambridge University Engineering Department., England, Dec. 2001.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Hierarchical Classifiers Approach for Emotions Recognition"
XIII Reunión de Trabajo en Procesamiento de la Información y Control (RPIC 2009), sep, 2009.