Available online at www.sciencedirect.com

ScienceDirect

Computer Speech and Language 25 (2011) 556–570

COMPUTER SPEECH AND LANGUAGE

www.elsevier.com/locate/csl

Spoken emotion recognition using hierarchical classifiers☆

Enrique M. Albornoz[a,b,*], Diego H. Milone[a,b], Hugo L. Rufiner[a,b,c]

[a] *Centro de I+D en Señales, Sistemas e INteligencia Computacional (SINC(i)), Fac. de Ingeniería y Cs. Hídricas, Univ. Nacional del Litoral, Santa Fe, Argentina*

[b] *Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET), Argentina*

[c] *Laboratorio de Cibernética, Facultad de Ingeniería, Univ. Nacional de Entre Ríos, Entre Ríos, Argentina*

Received 13 April 2010; received in revised form 1 October 2010; accepted 6 October 2010
Available online 26 October 2010

Abstract

The recognition of the emotional state of speakers is a multi-disciplinary research area that has received great interest over the last years. One of the most important goals is to improve the voice-based human–machine interactions. Several works on this domain use the prosodic features or the spectrum characteristics of speech signal, with neural networks, Gaussian mixtures and other standard classifiers. Usually, there is no acoustic interpretation of types of errors in the results. In this paper, the spectral characteristics of emotional signals are used in order to group emotions based on acoustic rather than psychological considerations. Standard classifiers based on Gaussian Mixture Models, Hidden Markov Models and Multilayer Perceptron are tested. These classifiers have been evaluated with different configurations and input features, in order to design a new hierarchical method for emotion classification. The proposed multiple feature hierarchical method for seven emotions, based on spectral and prosodic information, improves the performance over the standard classifiers and the fixed features.

Keywords: Emotion recognition; Spectral information; Hierarchical classifiers; Hidden Markov Model; Multilayer Perceptron

1. Introduction

In human interactions there are many ways in which information is exchanged (speech, body language, facial expressions, etc.). A speech message in which people express ideas or communicate has a lot of information that is interpreted implicitly. This information may be expressed or perceived in the intonation, volume and speed of the voice and in the emotional state of people, among others. The speaker's emotional state is closely related to this information, and this motivates its study. Two antagonistic ideas on the origin of emotions exist. One of these explains emotions from evolutionary psychology and the other as socially constructed (Prinz, 2004). The second theory claims that emotions are generated by society, and they find a different support in each culture. In evolutionary theory, it is widely accepted the "basic" term to define some universal emotions. The most popular set of basic emotions is the *big six*: happiness

☆ This paper has been recommended for acceptance by Dr. Zheng-Hua Tan.

* Corresponding author at: Univ. Nacional del Litoral CC 217, C. Universitaria, Paraje El Pozo, S3000 Santa Fe, Argentina. Tel.: +54 342 457 5233/39/44/45x191; fax: +54 342 457 5224.

E-mail addresses: emalbornoz@fich.unl.edu.ar (E.M. Albornoz), d.milone@ieee.org (D.H. Milone), lrufiner@fich.unl.edu.ar (H.L. Rufiner).

0885-2308/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.csl.2010.10.001

(joy), anger, fear, boredom, sadness, disgust and neutral. Ekman et al. (1969) researched it to argue in favour of emotion innateness and universality.

Over the last years the recognition of emotions has become a multi-disciplinary research area that has received great interest. This plays an important role in the improvement of human–machine interaction. Automatic recognition of speaker emotional state aims to achieve a more natural interaction between humans and machines. Also, it could be used to make the computer act according to the actual human emotion. This is useful in various real life applications as systems for real-life emotion detection using a corpus of agent-client spoken dialogues from a medical emergency call centre (Devillers and Vidrascu, 2007), detection of the emotional manifestation of fear in abnormal situations for a security application (Clavel et al., 2008), support of semi-automatic diagnosis of psychiatric diseases (Tacconi et al., 2008) and detection of emotional attitudes from child in spontaneous dialog interactions with computer characters (Yildirim et al., 2011). On the other hand, considering the other part of a communication system, progress was made in the context of speech synthesis too (Murray and Arnott, 2008).

The use of biosignals (such as ECG, EEG, etc.), face and body images are an interesting alternative to detect emotional states (Kim and André, 2008; Schindler et al., 2008; Vinhas et al., 2009). However, methods to record and use these signals are more invasive, complex and impossible in certain real applications. Therefore, the use of speech signals clearly becomes a more feasible option. Most of the previous works on emotion recognition have been based on the analysis of speech prosodic features and spectral information (Dellaert et al., 1996; Noguerias et al., 2001; Borchert and Dusterhoft, 2005; Luengo Gil et al., 2005; Batliner et al., 2011). Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Multilayer Perceptron (MLP) and several other one-level standard techniques have been explored for the classifier (Iliev et al., 2010; Albornoz et al., 2008; Lin and Wei, 2005; El Ayadi et al., 2007; Rong et al., 2007). Good results are obtained by standard classifiers but their performance improvement could have reached a limit. Fusion, combination and ensemble of classifiers could represent a new step towards better emotion recognition systems.

In last years some works using a combination of standard methods have been presented. A fusion scheme where a combination of results at the decision-level based on the outputs of separate classifiers (trained with different types of features) is proposed in Kim (2007). In Truong and van Leeuwen (2007), a similar idea in order to distinguish between laughter and speech is proposed. In this work, two ways to combine classifier outputs are presented: a linear combination of the outputs of independent classifiers and a second-level classifier trained with the outputs from a fixed set of independent classifiers. Two classification methods (stacked generalization and unweighted vote) were applied to emotion recognition of 6 emotional classes in Morrison et al. (2007). These classifiers improved modestly the performance of traditional classification methods, with recognition rates of 73.29% and 72.30%, respectively. In Schuller et al. (2004), a multiple stage classifier with support vector machine (SVM) is presented. Two-class decisions are repetitively made until only one class remains and hardly separable classes are divided at last. Authors built this partition based on expert knowledge or derived it from confusion matrices of a multiclass SVM approach. They reported an accuracy of 81.19% with 7 emotional classes. A two-stage classifier for five emotions is proposed in Fu et al. (2008) and the recognition rate reaches 76.1%. 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 Lee et al. (2009), Bayesian logistic regression and SVM classifiers in a binary decision tree are used. They reported 48.37% of unweighted recall on 5 emotional classes. The order of the classification at each layer of binary classification is motivated by appraisal theory of emotions (Lazarus, 2001). A binary multi-stage classifier guided by the dimensional emotion model is proposed in Xiao et al. (2009). They used six emotion states from the Berlin dataset and reported a classification rate of 68.60%. A true comparison among the results of all the previously mentioned methods is very difficult because they have used different corpus, training/test partitions, etc. Therefore, none of these results can be a baseline for direct comparison with our work and our own baselines are proposed. This will be discussed later.

A simple analysis of the output probability distribution of the HMM states obtained for different emotions is made in Wagner et al. (2007). However, the reasons for success and failure in confusion matrices are not usually analyzed. For example, in Schuller et al. (2004) and Fu et al. (2008) clustering was done based on confusion matrices of standard classifiers, expert knowledge or the goodness of SVM. In the present work, an analysis of spectral features is made in order to characterize emotions and to define groups. Emotions are grouped based on their acoustical features and a hierarchical classifier is designed. The emotions which are acoustically more similar agree with the emotions that are the most difficult to distinguish, as it can be seen in the confusion matrices reported in previous works (Noguerias et al., 2001; Albornoz et al., 2008; Borchert and Dusterhoft, 2005; El Ayadi et al., 2007). The proposed classifier is

Table 1
Corpus utterances grouped by emotion class.

	Emotion class						
	Anger	Boredom	Disgust	Fear	Joy	Sadness	Neutral
No. utterances	127	81	46	69	71	62	79

evaluated in the same experimental condition as standard classifiers showing important improvements in the recognition rates.

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

2. Acoustic analysis of emotions

2.1. Emotional speech corpus

As emotional expressions in real conversations are very changeable, the present goal is to achieve emotion recognition from spontaneous speech. On the other hand, the development of spontaneous-speech datasets is very expensive and they are commonly restricted. Although acted emotional expressions may not sound like real expressions, using it is an interesting approach. However, these become more useful if people whom develop them are not actors and the dataset naturalness is judged by expert listeners. In this context, the emotional speech signals used here were taken from an emotional speech database, developed at the Communication Science Institute of Berlin Technical University (Burkhardt et al., 2005). This corpus is a well-known acted database and it was used in several studies (Borchert and Dusterhoft, 2005; El Ayadi et al., 2007; Schuller et al., 2008; Yang and Lugger, 2010).[1] The corpus, consisting of 535 utterances, includes sentences performed under 6 plain emotions, and sentences in neutral emotional state. This corpus covers the big six emotions set except for boredom instead of surprise (Table 1 shows their distribution).

The same sentences were recorded in German by 10 actors, 5 females and 5 males, 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 s. 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 individuals was carried out to ensure the emotional quality and naturalness of the utterances, and the most confusing[2] utterances were eliminated (Burkhardt et al., 2005).

2.2. Acoustic analysis

The psychological conceptualization of affects, with two-dimensional and three-dimensional models, is widely known in the categorization of emotions (Kim, 2007; Cowie and Cornelius, 2003; Scherer, 2005). These models are often used to group emotions in order to define classes, such as those associated with low arousal and low pleasure versus those associated with high arousal and high pleasure. For example, in Kim and André (2008), a dyadic multi-level classifier based on this two-dimensional emotion model is presented.

In our study emotions are characterized both by spectral and prosodic information. How to take advantage from this acoustic evidence in the classification was studied, without taking into account information from the psychological level or the traditional taxonomy of human emotions.

[1] The corpus is freely accessible at http://pascal.kgw.tu-berlin.de/emodb/.

[2] The utterances were deleted when recognition errors were more than 20% and were judged as no natural by more than 40% of the listeners.

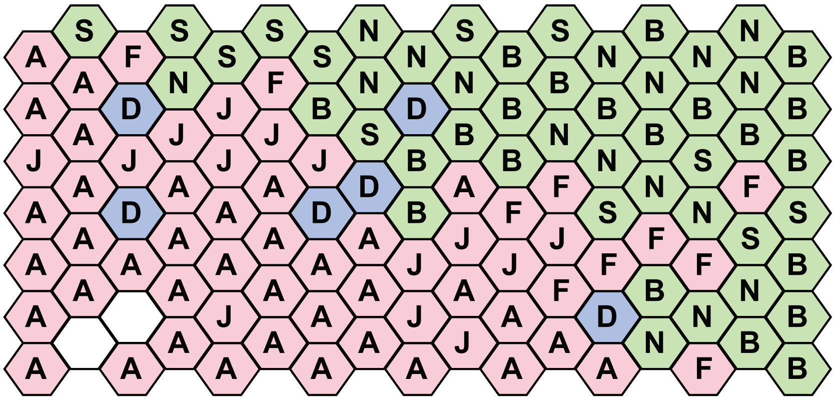

Fig. 1. Clustering of emotions using SOM (30 MLS coefficients).

The mean of the log-spectrum (MLS) on each frequency band along the frames was calculated for every utterance. Then, the average of the mean log-spectrums (AMLS) over all the utterances with same emotion were computed

$$S_\ell(k) = \frac{1}{N_\ell}\sum_{i=1}^{N_\ell}\frac{1}{N_i}\sum_{n=1}^{N_i}\log|v_{i\ell}(n,k)|, \tag{1}$$

where k is a frequency band, N_ℓ is the number of sentences for the emotion class ℓ, N_i is the number of frames in the utterance i and $v_{i\ell}(n, k)$ is the discrete Fourier transform of the signal i in the frame n.

The aim was to discover acoustical similarities among emotions using some underlying structure of the data. The main objective was to group emotions using significant features from the input data found in an unsupervised way. A very useful method to perform this task is the Self-Organizing Map (SOM). A SOM is a type of artificial neural network that is trained using unsupervised learning. It maps p-dimensional input patterns to a q-dimensional discretized map (Kohonen, 1995). It generally consists of an input layer and an output layer with feedforward connections from input to output and lateral connections among neurons in the output layer. SOM preserves the topological properties of the input space. Then nearby input patterns will be mapped preserving neighborhood relations. Nodes that are "close" together are going to interact differently than nodes that are "far" apart.

For each utterance, the MLS was used as an input pattern. With the aim of eliminating less informative coefficients, SOM with different number of MLS coefficients were trained. Tests with 200 (0–8000 Hz), 50 (0–2000 Hz), 40 (0–1600 Hz), 30 (0–1200 Hz) and 20 (0–800 Hz) coefficients were performed. The best clustering was obtained with the first 30 coefficients. Using the SOM as a classifier (after the training phase), every cell was labelled with the emotion that appeared most frequently at this place. The data projection which was obtained showed that certain emotion classes could be considered as groups when using the spectral features (Fig. 1). Data corresponding to Joy and Anger are displayed from the left-bottom corner to the centre of the map; whereas data on Boredom, Neutral and Sadness appear propagated from the right-top corner to the centre. On the other hand, Fear and Disgust patterns are placed in a more distributed manner. This visual information provided by the SOM map could be considered in order to group emotions in different ways. Thus, for example, a group could contain Joy, Anger and Fear emotions (JAF) whereas another contains Boredom, Neutral and Sadness emotions (BNS) and finally, Disgust emotion stands alone in a third group. Furthermore, a two-group schema where Disgust is placed in the JAFD group (JAF + Disgust), being the other group BNS could be considered.

Similar experiments using combination of MLS, mel frequency cepstral coefficients (MFCCs) and prosodic information were performed and results showed a similar clustering topology.

In order to validate the grouping approach previously presented with SOM, the AMLS features were visually explored. The most important information to discriminate among emotion classes was found between 0 and 1200 Hz. Fig. 2 shows this information for each emotional class. As it can be seen in the figures, emotions in the same SOM cluster are spectrally similar. For example, a similar shape and a maximum between 240 and 280 Hz in Joy, Anger and Fear can be noticed. A minimum is present close to 75 Hz in Joy, Anger, Fear and Disgust. On their part, Boredom, Neutral and Sadness have similar shape and a peak between 115 and 160 Hz.

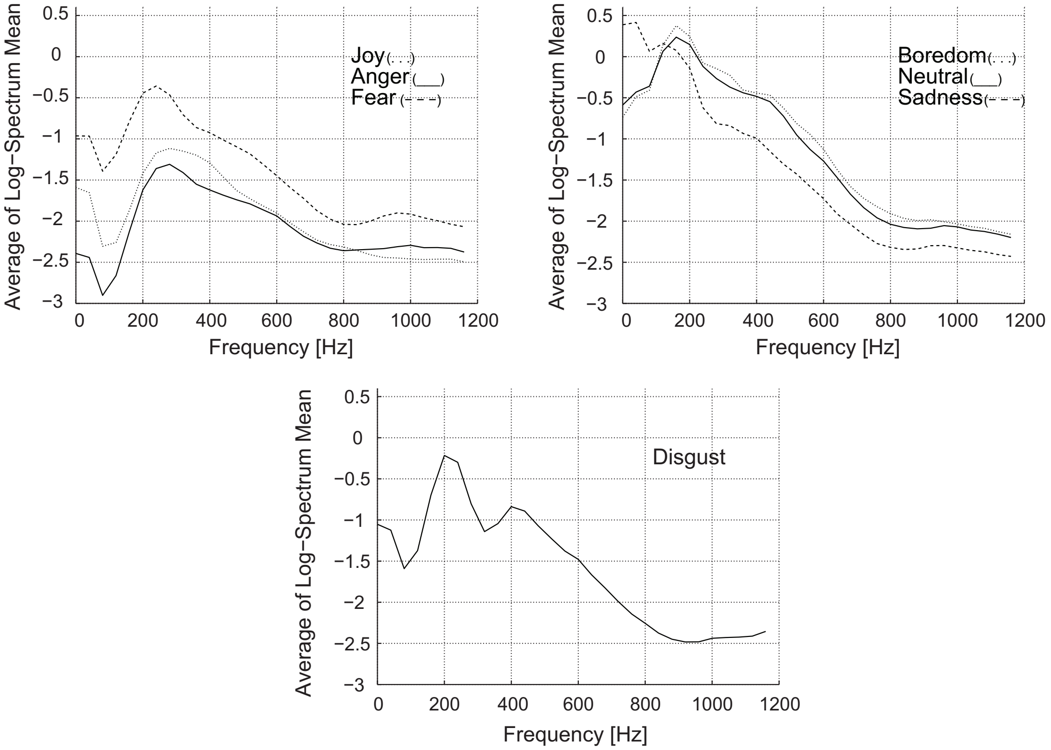

Fig. 2. Average of Mean Log-Spectrum for all emotion classes grouped by their spectral similarities.

Therefore, groups can be defined by using this spectral and prosodic information, independently from psychological considerations. Here the heuristic approach presented with AMLS is used in order to validate the information found previously with SOM.

This relevant knowledge for emotion grouping elicited from unsupervised clustering is used in the next section to design a hierarchical classifier.

3. Proposed method

The application of neural networks and statistical models to emotion classification are not so novel in emotion recognition. Moreover, as cited in Section 1, some combination of classifiers also were proposed in previous works. However, in that works the grouping of emotions was based on expert knowledge or psychological classifications. Unlike previous works, here we propose an unsupervised clustering based on acoustic-prosodic features in order to define a hierarchical classifier. Then, every part of this classifier is an independently defined unit, each with its own feature extraction and classification method.

It is clear from Table 1 that the distribution of emotions is unbalanced, where *anger* dominated the set (24% of the set). This characteristic of the database can bias the validation of methods. It is interesting to point out that almost all previous works did not address this issue, producing biased and non-comparable results. Unbalanced training datasets lead to unsuitable results in classifiers like MLP. Therefore, to avoid this problem, the dataset was balanced by equalizing the size of the classes which was performed by selecting randomly the same number of samples for all classes in each partition ($46 \times 7 = 322$ utterances). The transcriptions of the utterances are not considered and each utterance has one label that refers to the emotion expressed. Then, each utterance is an unique training or test pattern in a partition according to the random process which generates the partition.

Relying on some gender variability in emotional speech signals, other approaches were based on a previous stage for gender classification. However, in our proposal this is not contemplated. Indeed, in our approach the feature extraction

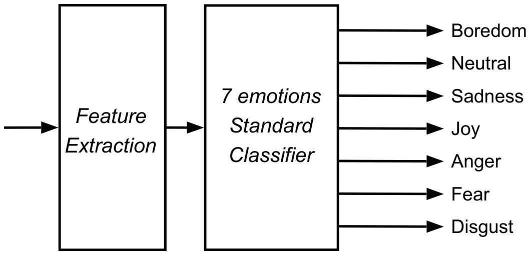

Fig. 3. Structure of standard one-level classifier for 7 emotions.

and classifier is aimed to find a recognizer that will be able to handle gender differences implicitly, that is, no specific blocks will be included for gender discrimination.

3.1. Features extraction and classification methods

For every emotional utterance, three kinds of characteristics were extracted: MLS, MFCCs and prosodic features. The MLS were computed as defined in Section 2.2. The spectrograms and the MFCC parametrization were calculated using Hamming windows of 25 ms with a 10 ms frame shift. The first 12 MFCC plus the first and second derivatives were also extracted using the *Hidden Markov Toolkit* (Young et al., 2001).

The use of prosodic features in emotion recognition has been discussed extensively. Often, in these works the classic methods to calculate *Energy* and F_0 along the signals were used (Deller et al., 1993). Many parameters can be extracted from prosodic features; usually the minimum, mean, maximum and standard deviation over the whole utterances were used. This set of parameters has already been studied and some works reported an important information gain to discriminate emotions (Borchert and Dusterhoft, 2005; Schuller et al., 2004; Adell Mercado et al., 2005).

In our work, combinations of features (MLS, the mean of every MFCCs and prosodic information) were arranged in vectors. In the previous SOM test, relevant information showed a similar structure to that observed in Fig. 1. For MLP tests, each dimension of the vector was normalized independently (from the rest of the remaining elements of that vector) by the maximum value that can be found for that dimension in the vector set.

In this work, some standard one-level classifiers (Fig. 3) are used as baseline reference. Classifiers are based on well-known techniques: MLP, GMM and HMM. MLP is a class of artificial neural network and it consists of a set of process units (simple perceptrons) arranged in layers. In the MLP, 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 (Haykin, 1998). The output of the neuron is the weighted sum of the inputs plus the bias term, and its activation is a function (linear or nonlinear) as

$$y = \mathcal{F}\left(\sum_{i=1}^{n} \omega_i x_i + \theta\right). \tag{2}$$

On the side of statistical theories, probabilistic distributions are used for classification. In this way, single Gaussian distribution is often used because it has important analytical properties, although they present limitations to model multimodal data. Superposition of multiple distributions would fit better for real data like speech features (typically multivariate). The *mixture of Gaussians* is a superposition formed as a finite linear combination of simple Gaussian densities and it is widely used in statistical pattern recognition (Bishop, 2006). For a n-dimensional data vector **x**, density could be modelled by a mixture of K Gaussians as

$$p(\mathbf{x}) = \sum_{k=1}^{K} \omega_k \mathcal{N}(\mathbf{x}|\mu_k, \Sigma_k), \tag{3}$$

where $\mathcal{N}$ is a single normal density defined by the mean vector (μ_k) and the covariance matrix (Σ_k). The mixing coefficients verify $\sum_k \omega_k = 1$ and $0 \leq \omega_k \leq 1$ for all k. By using a sufficient number of Gaussians, and by adjusting their means and covariances as well as the coefficients in the linear combination, almost any continuous density can be approximated to arbitrary accuracy (Bishop, 2006).

The HMMs are basically statistical models that describe sequences of events and it is a wide-used technique in speech and emotion recognition. In classification tasks, a model is estimated for every signal class. Usually in emotion recognition one class refers to a specific emotion. Thus, it would take into account as many models as signal classes to be recognized. 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 unknown signal (Rabiner and Juang, 1993). Here, the problem is presented as

$$\hat{E} = arg \max_{\ell} P(E_{\ell}|A), \tag{4}$$

where A is the sequence of acoustic features taken from speech signal and E_{ℓ} represent the models of the emotion ℓ.

In order to apply HMM and MLP, the *Hidden Markov Toolkit* (Young et al., 2001) and *Stuttgart Neural Network Simulator* (Zell et al., 1998) were used, respectively.

In a *design phase*, based on previous studies (Albornoz et al., 2008), the GMM and a two-state HMM were chosen. Tests increasing the number of Gaussian components in the mixtures were performed to find the optimal structure. In order to optimize the MLP performance, different numbers of inputs and neurons in the hidden layer were tested.

The estimation of recognition rate can be biased if only one training and one test partition is used. To avoid these estimation biases, a cross-validation with the leave-k-out method was performed (Michie et al., 1994). After the design phase, ten data partitions were generated for the *test phase*.

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.[3] The MLP training was stopped when the network reached the generalization peak with test data (Haykin, 1998). In HMM cases, the 20% used for tests was added to the standard train set.

3.2. Hierarchical classifier

In this section, a new multiple feature and hierarchical classification method based on the acoustic analysis described above is presented. The main motivation for the development of a hierarchical classifier is taking advantage of spectral emotion similarities to improve the emotion recognition rate. We also used the fact that better results can be achieved when the number of emotions decrease for the same standard classifier. Furthermore, the main differences between specific emotions are more evident with a particular feature vector and the best classification is obtained through a specialized classifier and structure. As it can be seen in Fig. 4, two hierarchical classifiers are proposed in two stages. Each stage is formed by one or two blocks, marked with dotted red lines, that contain a *feature extraction* section and a *classification* section. The classifier in Fig. 4(a) has a Stage I where the emotion utterance would be classified as belonging to one of the 3 emotional groups (BNS, JAF or Disgust), then it would be classified again in its corresponding block group (if it is not Disgust) and finally the emotion label is obtained. The second model (Fig. 4(b)) has a Stage I to classify the emotion utterance into one of the 2 emotional groups (BNS or JAFD) and a Stage II where the emotion label is obtained after the classification into the corresponding block was done.

4. Experiments and results

In this section we describe the details of the experiments performed in this work. We first discuss how we have chosen the structures of our hierarchical models in a design phase. Then we validate these models on a test phase.

To define the hierarchical model structure in each block, several configurations of MLP and HMM with different parameter vectors were evaluated. An extra data partition was extracted to evaluate every feasible block of the hierarchical model structure in the *design phase*. As the other partitions, each emotional category has 46 utterances. Finally, definitive model stages were chosen and assembled with classifiers that achieved best results in isolated block tests, with the design partition.

In every MLP block test, 15 feature vectors (FV) were tested in 3 different hidden layer configurations (90, 120 and 150 perceptrons). Table 2 shows the number of characteristics for each vector and the kinds of features it includes. The

[3] In MLP experiments each partition has 196 utterances for training, 63 utterances for generalization and 63 utterances for validation.

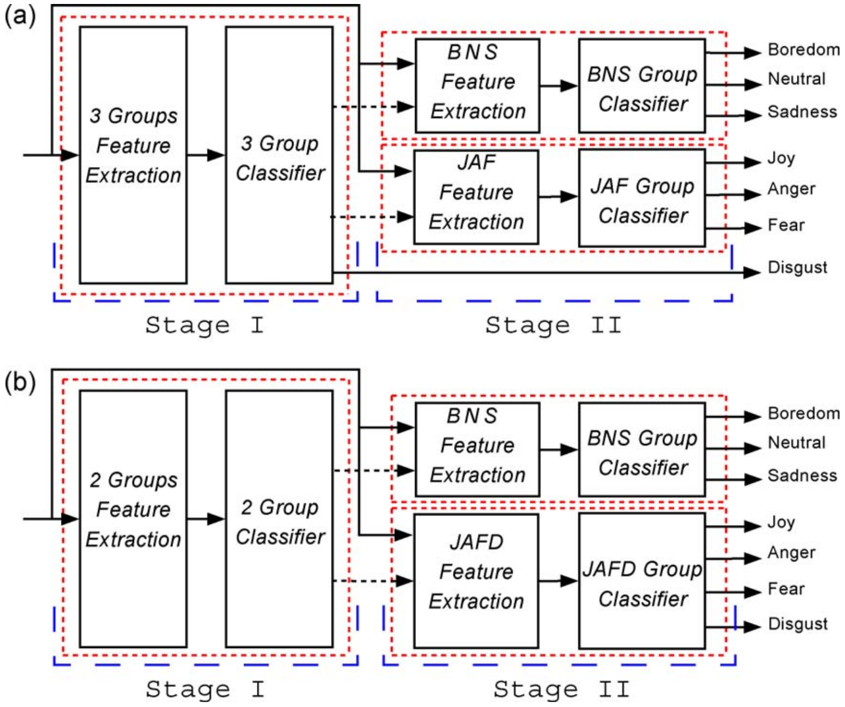

Fig. 4. General structure of the two hierarchical classifier for 7 emotions: (a) classifier for 3 emotional groups in Stage I; (b) classifier for 2 emotional groups in Stage I.

MFCC and Log-Spectrum coefficients were computed in frames, and then their means were calculated over all frames. For example, the feature vector *FV14* includes 12 mean MFCC, the F_0 mean and the Energy mean. On the other hand, a 36 coefficient frame vector was used for HMM tests (12 MFCCs plus delta and acceleration), as in Albornoz et al. (2008).

A comparative analysis between GMM and HMM for recognition of seven emotions was presented in Albornoz et al. (2008). The best results for 7 emotions were achieved using a two state HMM with mixtures of 30 Gaussians, using a MFCC parametrization with delta and acceleration coefficients, whereas the best result with GMM for 7 emotions was with mixtures of 32 Gaussians. Here, the same systems with ten balanced partitions were tested with cross-validation in order to obtain baseline results. The classification rate was 63.49% with GMM and 68.57% with HMM. In our study, the best performance from a MLP with a number of output nodes equal to seven emotions was 66.83% with cross-validation. This network, considered here as a baseline, was composed of 90 hidden neurons using *FV46* as input.

4.1. Design phase

The evaluation of every block with several configurations in an isolated way is proposed here. As already mentioned, one particular data partition (*design partition*) was extracted in order to define the structure of the hierarchical multiple

Table 2
Feature vectors used in MLP tests.

Parameters	FV12	FV14	FV16	FV18	FV20	FV30	FV32	FV34	FV36	FV38	FV42	FV44	FV46	FV48	FV50
12 Mean 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 3
Performance of MLP for 3 emotional groups (*design part.*). 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

feature system. In Fig. 4, it is possible to identify 5 different blocks to be tested (two in Stage I and three in the Stage II). Full experiments with MLP and HMM for every block were done. For Stage I in the hierarchical classifier, six different options were evaluated: (a) to re-group HMM baseline outputs into 3 emotional groups ($\mathrm{HMM}^7\mathrm{g}^3$); (b) to model each of the 3 emotional groups with one HMM (HMM^3); (c) to use a MLP with 3 output neurons (MLP^3); (d) to re-group HMM baseline outputs into 2 emotional groups ($\mathrm{HMM}^7\mathrm{g}^2$); (e) to model each of the 2 emotional groups with one HMM (HMM^2); and (f) to use a MLP with 2 output neurons (MLP^2). Options (a) and (d) were computed using configurations for seven emotions that achieved best results in Albornoz et al. (2008). Using these best configurations for HMM from previous work, the number of Gaussian components in the mixture were altered in order to find the best model for (b) and (e). The best models have 30 Gaussians for 3 emotional groups and 8 Gaussians for 2 emotional groups. Tables 3 and 4 show the MLP results for each feature vector with train and validation data, for 3 and 2 emotional groups respectively. The *Best Net* column shows the settings that worked better and these are presented with the number of neurons in the input, hidden and output layers (as *Input* + *Hidden* + *Output*). The best results obtained for Stage I are summarized in Tables 5 and 6. Table 5 shows that MLP achieved the best result for 3 emotional groups but it is the worst classifying Disgust. This could be because MLP is not a good classifier when the classes are noticeably unbalanced, as in Stage I for 3 emotional groups. Furthermore, the MLP reached a performance of 100% for the 2 almost balanced groups (Table 6).

Table 4
Performance of MLP for 2 emotional groups (*design part.*). Classification rate in [%].

Input	Best net	Train	Validation
FV12	12+90+2	98.47	98.41
FV14	14+90+2	92.35	98.41
FV16	16+90+2	93.88	98.41
FV18	18+90+2	95.41	92.06
FV20	20+90+2	93.37	92.06
FV30	30+150+2	100.00	95.24
FV32	32+90+2	100.00	95.24
FV34	34+90+2	97.45	95.24
FV36	36+120+2	93.37	90.48
FV38	38+90+2	98.98	93.65
FV42	42+120+2	100.00	98.41
FV44	44+90+2	100.00	96.83
FV46	46+90+2	**98.98**	**100.00**
FV48	48+120+2	96.43	98.41
FV50	50+150+2	100.00	96.83

Table 5
Performance of classification models for 3 emotional groups in Stage I (*design part.*). Classification rate in [%].

	$\mathrm{HMM}^7\mathrm{g}^3$	HMM^3	MLP^3
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 6
Performance of classification models for 2 emotional groups in Stage I (*design part.*). Classification rate in [%].

	$\mathrm{HMM}^7\mathrm{g}^2$	HMM^2	MLP^2
JAFD	94.44	88.89	**100.00**
BNS	85.19	96.30	**100.00**
Average	90.48	92.06	**100.00**

Table 7
Best performances for isolated Stage II classification (*design part.*). Classification rate in [%].

Emotional group	Stage II model	Performance
JAF	MLP (46+90+3)	**85.19**
	HMM (26 Gauss.)	74.07
JAFD	MLP (12+90+4)	66.67
	HMM (20 Gauss.)	**77.78**
BNS	MLP (44+150+3)	**81.48**
	HMM (4 Gauss.)	77.78

For each feasible block in Stages II of both proposed models, HMM and MLP tests were done using *design partition* 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. The best results for HMM were 74.07% for JAF test with 26 Gaussians in the mixtures, 77.78% for JAFD test with 20 Gaussians in the mixtures, while only 4 Gaussians achieved 77.78% for the BNS case. MLP experiments using every FV and three different network structures were done. The best results for the isolated blocks of Stage II are shown in Table 7.

The results presented show that it is very important to deal with particular problems using specific FV in order to achieve a better performance. A similar analysis could justify the choice of classifiers and their structure in each block. The best configurations for every block are used in the next section in order to validate the method.

4.2. Test phase

Ten partitions were generated in order to validate the hierarchical multiple feature system. These partitions were extracted from the whole corpus at random as it was done to produce the design partition.[4] Block structures and input features were selected from those configurations that achieved the best results in the *design phase*.

The best results obtained for three and two groups of emotions in Stage I are summarized in Table 8.

The overall classifier performance using 3 emotional groups in Stage I is presented in the “Best” column of Table 9. In the second column, Disgust scores are displayed as result of Stage I for each model. Patterns classified as JAF in Stage I are evaluated with both models in Stage II in order to identify the specific emotion. The recognition averages for these three emotions are shown in the third and fourth columns of the table, for each classifier in the second stage. The same information can be seen for BNS in the fifth and sixth columns. The *Best* column shows the performance

[4] Here, the *design partition* is not used.

Table 8
Best performances for isolated Stage I classification. Classification rate in [%].

Stage I	Model	Performance
3 emotional groups	HMM^{7}g^3 (30 Gaussians)	**89.84**
	HMM^3 (30 Gaussians)	86.82
	MLP (34+120+3)	82.06
2 emotional groups	HMM^{7}g^2 (30 Gaussians)	92.86
	HMM^2 (8 Gaussians)	90.16
	MLP (46+90+2)	**93.02**

Table 9
Final performances of hierarchical model for 3 emotional groups in Stage I. Classification rate in [%].

Stage I		Stage II				Best
Model	Disgust	JAF		BNS		
		HMM	MLP	HMM	MLP	
HMM^{7}g^3	**80.00**	63.70	**71.48**	**69.26**	62.96	**71.75**
HMM^3	68.89	59.63	**67.78**	**71.48**	62.22	**69.52**
MLP	57.78	56.30	**64.07**	**68.89**	62.22	65.24

computed by a combination of best models for JAF and BNS groups. The classification rate is calculated as proportional to the test patterns in each emotion group ($R = (R_D + 3R_{JAF} + 3R_{BNS})/7$). As it was previously mentioned, the number of test patterns is balanced. As can be observed in this table, MLPs are always better than HMMs for JAF block whereas HMMs obtained a better performance for BNS block. This could be related with the fact that MLP only performs a static classification, while HMMs take advantage of additional temporal information in order to produce better discrimination within BNS emotional group.

Therefore, considering 3 emotional groups in Stage I, the best multiple feature hierarchical model is formed by HMM re-grouped (HMM^{7}g^3) with 30 Gaussians in mixtures in Stage I; MLP with FV46 and 90 hidden neurons for the JAF block and HMM with 4 Gaussians in mixtures for the BNS block (Fig. 5).

Table 10 shows the performance for JAFD and BNS blocks with both models, for each model in Stage I. The performance for the best combination considering each model for 2 emotional groups in Stage I is: 66.99% for HMMs re-grouped (HMM^{7}g^2), 64.44% for 2 HMMs (HMM^2) and 66.03% for MLP. Considering 2 emotional groups in Stage I, the best multiple feature hierarchical model is formed by a HMM re-grouped (HMM^{7}g^2) with 30 Gaussians in mixtures in the Stage I; a HMM with 20 Gaussians in mixtures for the JAFD block and a HMM with 4 Gaussians in mixtures for the BNS block. In this model, the HMMs obtained a better performance for BNS block again, whereas the HMMs are better than MLP for the JAFD block. For this schema, there was no configuration that improves the baseline performance.

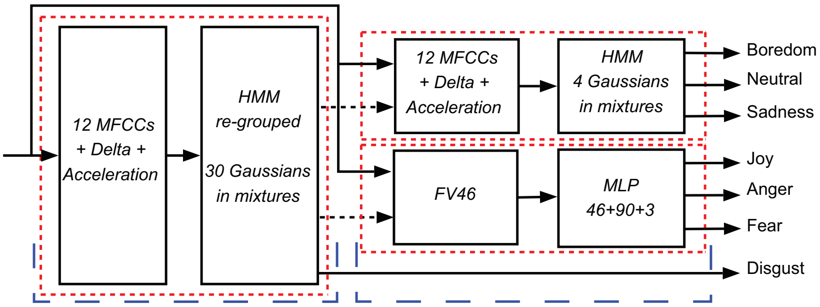

Fig. 5. Structure of the best hierarchical classifier for 3 emotional groups in Stage I.

Table 10
Final performance of hierarchical model for 2 emotional groups in Stage I. Classification rate in [%].

Stage I	Stage II				Best
Model	JAFD		BNS		
	HMM	MLP	HMM	MLP	
HMM^{7}g^2	**65.28**	58.61	**69.26**	62.96	66.99
HMM^2	**57.78**	55.56	**73.33**	64.82	64.44
MLP	**63.33**	60.00	**69.63**	63.33	66.03

Table 11
Overall performance of standard classifiers vs. hierarchical model. Classification rate in [%].

Model	Performance
GMM	63.49
MLP	66.83
HMM	68.57
Hierarchical 2	66.99
Hierarchical 3	**71.75**

Table 12
Confusion matrix of standard one-level HMM.

Emotion	Joy	Anger	Fear	Disgust	Boredom	Neutral	Sadness
Joy	60	16	9	5			
Anger	14	**71**	2	1		2	
Fear	11	5	58	3	5	8	
Disgust	1	3	6	72	5	3	
Boredom			2	6	**55**	21	6
Neutral	2		4	5	27	51	1
Sadness			1	2	11	11	65

Table 11 contains a comparison between standard one-level classifiers and the best multi-feature hierarchical classifiers proposed here. Results show that hierarchical method improves the performance in 3.18% over the best standard classifier, with ten-fold cross-validation. Further analysis of these results can be obtained by confusion matrices. The confusion matrices give a good representation of results per each class allowing to make a detailed analysis of performance and to find the main classification errors. The confusion matrices (adding all partitions) of the standard HMM model (68.57%) and the Hierarchical 3 model that achieved the best performance (71.75%) are shown in Tables 12 and 13, respectively. The rows symbolize the actual class labels and the columns have the predicted labels of emotions, therefore, the main diagonal shows the emotions correctly recognized. It can be observed in Table 12 that the most important confusions are within the JAF and BNS emotional groups, as expected. In this table can be

Table 13
Confusion matrix of the Hierarchical 3 classifier using 3 emotional groups in Stage I.

Emotion	Joy	Anger	Fear	Disgust	Boredom	Neutral	Sadness
Joy	**63**	14	8	5			
Anger	17	65	5	1		2	
Fear	6	3	**65**	3	5	8	
Disgust	1	3	6	72	5	3	
Boredom			2	6	53	17	12
Neutral	2		4	5	15	**64**	
Sadness			1	2	10	7	**70**

observed the major confusions in the underlined numbers: between Boredom and Neutral there are 48 errors (27 + 21); there are 20 confusions between Joy and Fear and there are 30 errors between Joy and Anger. As the hierarchical model deals individually with each emotional group, it is able to identify better the emotions within each group. In Table 13, the confusions within emotional groups are shaded and it is possible to discern a lesser confusion in each of them, although as can be seen these remain difficult to discriminate. Obviously the confusions are not fully resolved, however, the proposed approach has reduced confusion within each emotional group.

Classifying groups of emotions using acoustical similarities allows to process the problem more efficiently. The configurations that improve the performance of one-level classifiers considered a schema with 3 emotional groups in Stage I. It can be observed in Table 9 that MLP achieved the worst result in Stage I mainly classifying Disgust. This could be because MLP is not a good classifier when the classes are unbalanced, as we mentioned. Contrarily, both HMM configurations obtained good results in Stage I. For the JAF block it is important to use all features to generate a space separable with a linear classifier. Modelling the temporal dynamics of MFCC with HMM is very useful in the case of BNS. It uses the most distinctive features, few Gaussians in the mixtures, and obtains the best results.

Results indicate that every group of emotions should be dealt with a specific model in order to improve the recognizer performance. For example, HMM are better in discriminating between B, N and S whereas that MLP is better classifying among J, A and F. In the same way, it was shown that some features are better to distinguish specific emotions. As already discussed for Stage I, MFCCs plus delta and acceleration coefficients perform better in discriminating 3 emotional groups while FV46 performs better in distinguishing 2 emotional groups.

5. Conclusions and future work

In this paper a characterization of emotions and analysis of their similarities based on the acoustical features were presented. Through the SOM unsupervised clustering and spectral analysis, new classes for emotion grouping have been proposed. Then, a new hierarchical method for emotion classification supported by such acoustic analysis was proposed. Two schemas based on blocks of acoustical groups were presented. Experiments with several feature vectors and internal structure for MLP were performed for each block. Also, tests increasing the number of Gaussians in mixtures for HMM were done.

Results show that the spectral information combined with prosody allow emotion grouping and it could guide the development of hierarchical classifiers. These models improve the recognition rates of standard one-stage classifiers. Furthermore, it was shown that prosody combined with spectral features improves the results in the emotion recognition task.

In future works the hierarchical classifier will be tested with noisy signals. Furthermore, these results will be compared with another model which take into account gender variability in an explicit manner. Similar analyses on other languages are also planned.

Acknowledgements

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

References

Adell Mercado, J., Bonafonte Cávez, A., Escudero Mancebo, D., 2005. Analysis of prosodic features: towards modelling of emotional and pragmatic attributes of speech. In: XXI Congreso de la Sociedad Española. Procesamiento del Lenguaje Natural – SEPLN 2005, no. 35, Granada, España, pp. 277–284.

Albornoz, E.M., Crolla, M.B., Milone, D.H., 2008. Recognition of emotions in speech. In: Proceedings of XXXIV Congreso Latinoamericano de Informática – CLEI, Santa Fe, Argentina, pp. 1120–1129.

Batliner, A., Steidl, S., Schuller, B., Seppi, D., Vogt, T., Wagner, J., Devillers, L., Vidrascu, L., Aharonson, V., Kessous, L., Amir, N., 2011. Whodunnit – searching for the most important feature types signalling emotion-related user states in speech. Computer Speech & Language 25 (1), 4–28.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning, 1st edition. Springer.

Borchert, M., Dusterhoft, A., 2005. Emotions in speech – experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments. In: Proceedings of IEEE International Conference on Natural Language Processing and Knowledge Engineering. IEEE NLP-KE 2005, pp. 147–151.

Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B., 2005. A database of German emotional speech. In: 9th European Conference on Speech Communication and Technology – Interspeech'2005, pp. 1517–1520.

Clavel, C., Vasilescu, I., Devillers, L., Richard, G., Ehrette, T., 2008. Fear-type emotion recognition for future audio-based surveillance systems. Speech Communication 50 (6), 487–503.

Cowie, R., Cornelius, R., 2003. Describing the emotional states that are expressed in speech. Speech Communication 40 (1–2), 5–32.

Dellaert, F., Polzin, T., Waibel, A., 1996. Recognizing emotions in speech. In: Proceedings of International Conference on Spoken Language Processing – ICSLP'96, vol. 3, Philadelphia, PA, pp. 1970–1973.

Deller, J.R., Proakis, J.G., Hansen Jr., J.H., 1993. Discrete-time Processing of Speech Signals. Prentice Hall PTR, Upper Saddle River, NJ, USA.

Devillers, L., Vidrascu, L., 2007. Speaker Classification II: Selected Projects, vol. 4441/2007 of Lecture Notes in Computer Science. Springer-Verlag, Berlin/Heidelberg, Chapter: Real-life emotion recognition in speech, pp. 34–42.

El Ayadi, M., Kamel, M., Karray, F., 2007. Speech emotion recognition using Gaussian mixture vector autoregressive models. In: IEEE International Conference on Acoustics, Speech and Signal Processing – ICASSP 2007, vol. 4, pp. IV-957–IV-960.

Ekman, P., Sorenson, E.R., Friesen, W.V., 1969. Pan-cultural elements in facial displays of emotions. Science 164 (3875), 86–88.

Fu, L., Mao, X., Chen, L., 2008. Speaker independent emotion recognition based on SVM/HMMs fusion system. In: International Conf. on Audio, Language and Image Processing. ICALIP 2008, pp. 61–65.

Haykin, S., 1998. Neural Networks: A Comprehensive Foundation, 2nd edition. Prentice Hall.

Iliev, A.I., Scordilis, M.S., Papa, J.P., Falcao, A.X., 2010. Spoken emotion recognition through optimum-path forest classification using glottal features. Computer Speech & Language 24 (3), 445–460.

Kim, J., 2007. Robust Speech Recognition and Understanding. I-Tech Education and Publishing, Vienna, Austria, Chapter: Bimodal emotion recognition using speech and physiological changes, pp. 265–280.

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

Kohonen, T., 1995. The Self-Organizing Map. Springer-Verlag.

Lazarus, R., 2001. Appraisal Processes in Emotion: Theory, Methods, Research (Series in Affective Science). Oxford University Press, USA, Chapter: Relational meaning and discrete emotions, pp. 37–67.

Lee, C., Mower, E., Busso, C., Lee, S., Narayanan, S., 2009. Emotion recognition using a hierarchical binary decision tree approach. In: Interspeech 2009, Brighton, UK, pp. 320–323.

Lin, Y.-L., Wei, G., 2005. Speech emotion recognition based on HMM and SVM. In: Proceedings of International Conference on Machine Learning and Cybernetics, vol. 8, pp. 4898–4901.

Luengo Gil, I., Navas Cordón, E., Hernáez Rioja, I.C., Sánchez de la Fuente, J., 2005. Reconocimiento automático de emociones utilizando parámetros prosódicos. Procesamiento del lenguaje natural 35, 13–20.

Michie, D., Spiegelhalter, D., Taylor, C., 1994. Machine Learning, Neural and Statistical Classification. Ellis Horwood, University College, London.

Morrison, D., Wang, R., Silva, L.C.D., 2007. Ensemble methods for spoken emotion recognition in call-centres. Speech Communication 49 (2), 98–112.

Murray, I.R., Arnott, J.L., 2008. Applying an analysis of acted vocal emotions to improve the simulation of synthetic speech. Computer Speech & Language 22 (2), 107–129.

Noguerias, A., Moreno, A., Bonafonte, A., Mariño, J., 2001. Speech emotion recognition using hidden Markov models. In: European Conference on Speech Communication and Technology – Eurospeech 2001, Aalborg, Denmark, pp. 2679–2682.

Prinz, J.J., 2004. Which emotions are basic? In: Evans, D., Cruse, P. (Eds.), Emotion, Evolution, and Rationality. Oxford University Press.

Rabiner, L., Juang, B.-H., 1993. Fundamentals of Speech Recognition. Prentice-Hall, Inc., Upper Saddle River, NJ, USA.

Rong, J., Chen, Y.-P.P., Chowdhury, M., Li, G., 2007. Acoustic features extraction for emotion recognition. In: 6th IEEE/ACIS International Conference on Computer and Information Science – ICIS 2007, pp. 419–424.

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

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

Schuller, B., Rigoll, G., Lang, M., 2004. Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture. In: Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing – ICASSP'04, pp. I-577–I-580.

Schuller, B., Vlasenko, B., Arsic, D., Rigoll, G., Wendemuth, A., 2008. Combining speech recognition and acoustic word emotion models for robust text-independent emotion recognition. In: IEEE International Conference on Multimedia and Expo'08, pp. 1333–1336.

Tacconi, D., Mayora, O., Lukowicz, P., Arnrich, B., Setz, C., Tröster, G., Haring, C., 2008. Activity and emotion recognition to support early diagnosis of psychiatric diseases. In: Proceedings of 2nd International Conference on Pervasive Computing Technologies for Healthcare'08, Tampere, Finland, pp. 100–102.

Truong, K.P., van Leeuwen, D.A., 2007. Automatic discrimination between laughter and speech. Speech Communication 49 (2), 144–158.

Vinhas, V., Reis, L.P., Oliveira, E., 2009. Dynamic multimedia content delivery based on real-time user emotions. Multichannel online biosignals towards adaptative GUI and content delivery. In: International Conference on Bio-inspired Systems and Signal Processing – Biosignals 2009, Porto, Portugal, pp. 299–304.

Wagner, J., Vogt, T., André, E., 2007. A systematic comparison of different HMM designs for emotion recognition from acted and spontaneous speech. In: Proceedings of the 2nd International Conference on Affective Computing and Intelligent Interaction – ACII'07. Springer-Verlag, Berlin/Heidelberg, pp. 114–125.

Xiao, Z., Dellandréa, E., Dou, W., Chen, L., 2009. Recognition of emotions in speech by a hierarchical approach. In: International Conference on Affective Computing and Intelligent Interaction (ACII), pp. 312–319.

Yang, B., Lugger, M., 2010. Emotion recognition from speech signals using new harmony features. Signal Processing 90 (5), 1415–1423 (special Section on Statistical Signal & Array Processing).

Yildirim, S., Narayanan, S., Potamianos, A., 2011. Detecting emotional state of a child in a conversational computer game. Computer Speech & Language 25 (1), 29–44.

Young, S., Evermann, G., Kershaw, D., Moore, G., Odell, J., Ollason, D., Valtchev, V., Woodland, P., 2001. The HTK Book (for HTK Version 3.1). Cambridge University Engineering Department, England.

Zell, A., Mamier, G., Vogt, M., Mache, N., Hubner, R., Doring, S., Herrmann, K.-U., Soyez, T., Schmalzl, M., Sommer, T., Hatzigeorgiou, A., Posselt, D., Schreiner, T., Kett, B., Clemente, G., 1998. SNNS (Stuttgart Neural Network Simulator) – user manual Version 4.2, 70565 Stuttgart, Fed. Rep. of Germany, sNNS User Manual Version 4.0.