Noname manuscript No.
(will be inserted by the editor)

Feature Extraction based on Bio-inspired Model for Robust Emotion Recognition

Enrique M. Albornoz · Diego H. Milone · Hugo L. Rufiner

Received: October 3, 2015 / Revised: December X, 2015 / Accepted: date

Abstract Emotional state identification is an important issue to achieve more natural speech interactive systems. Ideally, these systems should also be able to work in real environments in which generally exist some kind of noise. Several bio-inspired representations have been applied to artificial systems for speech processing under noise conditions. In this work, an auditory signal representation is used to obtain a novel bio-inspired set of features for emotional speech signals. These characteristics, together with other spectral and prosodic features, are used for emotion recognition under noise conditions. Neural models were trained as classifiers and results were compared to the well-known mel-frequency cepstral coefficients. Results show that using the proposed representations, it is possible to significantly improve the robustness of an emotion recognition system. The results were also validated in a speaker independent scheme and with two emotional speech corpora.

Keywords Robust Emotion Recognition · Auditory Representation · Multilayer Perceptron

1 Introduction

Emotions represent a very important part in human communications and they can be perceived in speech signals, in facial expressions, in body posture, in biosignal as electrocardiograph (ECG), blood pressure and electroencephalogram (EEG), among others. Affective state recognition has received much attention in recent years, mainly because its result could be useful in various applications as depression and suicide risk assessment using speech analysis (Cummins et al 2015), systems for real-life emotion detection in a medical emergency call centre (Devillers and Vidrascu 2007), automatic prediction of frustration (Kapoor et al 2007), detection 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), music recommendation (Chin et al 2014) and detection of emotional attitudes of a child in dialogue interactions with a computer (Yildirim et al 2011).

As was mentioned, emotions are expressed in more than one modality or communication channel and several studies explored multi-modal systems in the context of emotion recognition (Chanel et al 2009; Giakoumis et al 2013; Kim and André 2008; Schindler et al 2008; Shojaeilangari et al 2015; Wang et al 2012; Wöllmer et al 2013). In spite of good results, the methods to record and to use these signals are invasive, complex and impossible in certain real applications. Therefore, the use of speech signals is clearly the most feasible option if it is available.

There are many proposals to deal with the problem of feature extraction from speech which are focussed on different aspects: speech production, characteristics of speech signals, speech perception, etc. Consequently, the chosen approach will determine the features that could be extracted for source, system, prosody, etc. Excitation-source features have been explored for emotional speech signals (Iliev et al 2010; Koolagudi and Rao 2012). However, most of the re-

Enrique M. Albornoz · Diego H. Milone · Hugo L. Rufiner
Research Institute for Signals, Systems and Computational Intelligence, sinc(i), UNL-CONICET. Faculty of Engineering and Water Science, Ciudad Universitaria, Paraje El Pozo, (S3000) Santa Fe (Argentina).
Tel.: +54 (342) 457-5233 ext 191 - FAX: +54 (342) 4575224
E-mail: emalbornoz@sinc.unl.edu.ar,
E-mail: dmilone@sinc.unl.edu.ar,
E-mail: lrufiner@sinc.unl.edu.ar
`http://fich.unl.edu.ar/sinc`

searchers have addressed their attention to the analysis of speech prosodic features and spectral information (Borchert and Dusterhoft 2005; Dellaert et al 1996; Luengo Gil et al 2005; Noguerias et al 2001). Mel-frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCCs), perceptual linear prediction coefficients (PLPCs), and formant features are some of the widely known features used in the literature (Batliner et al 2011; El Ayadi et al 2011).

With regard to classification, several standard techniques have been explored for emotion recognition, among which we can mention Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Multilayer Perceptron (MLP), Support Vector Machines (SVM), k-nearest neighbour (k-NN), Bayesian classifiers (El Ayadi et al 2011, 2007; Lin and Wei 2005; Wagner et al 2007). Good results are obtained by standard classifiers but their performance improvement could have reached a limit. At present, the combination of standard methods have become the focus of state-of-the-art studies (El Ayadi et al 2011; Zeng et al 2009). Morrison et al (2007) applied stacked generalization and unweighted vote to emotion recognition. Kim (2007) proposed a fusion scheme, in which the outputs of separate classifiers are combined at the decision-level. A similar idea to distinguish between laughter and speech was proposed by Truong and van Leeuwen (2007). Schuller et al (2004) proposed a multiple stage classifier using SVM, and a two-stage classifier using SVM and HMM was proposed by Fu et al (2008). Bayesian logistic regression and SVM classifiers in a binary decision tree is proposed by Lee et al (2009), and its structure is motivated by the appraisal theory of emotions (Lazarus 2001). A hierarchical model and a binary multi-stage classifier guided by the dimensional emotion model were proposed by Lugger and Yang (2008) and Xiao et al (2009), respectively. A hierarchical classifier was developed using clusters of emotions defined from spectral and prosodic features, rather than psychological considerations (Albornoz et al 2011).

Some of current challenges are the coupling of tasks, continuous modelling, robustness, more realism, cross-corpus, and the ability to operate using never-seen languages (Schuller and Weninger 2012; Albornoz and Milone 2016). Robust emotion recognition is an issue that have not received sufficient attention and it is a current challenge (Koolagudi and Rao 2012; Schuller et al 2013; Zeng et al 2009). While a few works used noisy corpus taken from real environments (Sztahó et al 2011; Tawari and Trivedi 2010; Rao and Koolagudi 2013), there are some that explored the controlled noise addition (Han et al 2012; Kandali et al 2010; Pao et al 2007; Schuller et al 2007). Sztahó et al (2011) proposed different acoustical features and a SVM classifier for emotion classification using spontaneous speech databases. For that work, they developed a database with spontaneous telephone conversations played by actors, but there is not an objective evaluation of noise level. Tawari and Trivedi (2010) proposed a module for adaptive noise cancellation as a first stage in a speech emotion recogniser. They explored the performance for white Gaussian noise and noises in a car for different scenarios: highway, parking lot and city street. Furthermore, several classifiers were explored for emotional speech recognition in Mandarin (Pao et al 2007). In that work, an experiment with k-NN was performed to select the more representative features (among speech formants, MFCC, LPCC, LPC, RASTA-PLP, etc.) without noise. Experimental results, with white Gaussian noise, showed that GMM outperformed other classifiers in high presence of noise while a model based on a modified k-NN achieved highest accuracy for minor amounts of noise. Schuller et al (2007) have explored different noise conditions for two acted and one spontaneous databases. For each database, a set of 4000 acoustic features (AF) was extracted, and 25 extra supra-segmental prosodic features (based on experience and literature) were manually extracted for spontaneous database. Random forests was used as classifier. The effect of white noise addition, using all features and the best features computed with Sequential Forward Floating Search (SFFS), was explored for acted databases. Reduction in the feature set helped to improve the performance in most cases. Kandali et al (2010) used a GMM classifier in order to compare a feature set based on eigen values of autocorrelation matrix (EVAM) against MFCCs. An emotional speech database, containing six basic emotions and neutral, was created for five native languages of Assam. Emotion recognition was performed for each language and for a cross-language scheme with additive babble noise for 5 dB and 0 dB SNR. Han et al (2012) used prosodic and quality features with a MLP classifier to classify seven emotional states in Chinese language. After feature extraction stage, they proposed a stage that perform the Canonical Correlation Based on Compensation. They reported an improvement in the performance for robust classification under 10 dB SNR, however, the type of noise used was not pointed out. Eyben et al (2012) proposed a robust system to evaluate five additive noise types (babble, street, office, white and music) using three databases. They used a feature set of 4.368 audio features (provided for a challenge in Interspeech 2011) and a classifier based on SVM with linear kernel. The results were validated with three training methods: mismatched condition, matched condition and multi-

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Feature extraction based on bio-inspired model for robust emotion recognition"
Soft Computing, Vol. 21, No. 17, pp. 5145-5158, Sep. 2017.

condition. The ComParE feature set[1] has been evaluated for affect recognition in real-life acoustic conditions (Eyben et al 2013). Rao and Koolagudi (2013) explored spectral and prosodic features for performing robust emotion recognition. This book focuses on conventional and non-conventional features to discriminate emotions in noisy and emotional environments. However, it does not present any feature set bio-inspired in auditory models.

In the present work, we propose a novel set of features based on a bio-inspired model for emotion recognition. In order to compute these features, the auditory model proposed by Shamma et al (1986) was used. This model tries to mimic the auditory system, then it is interesting to know if the properties provided by the model are useful for the discrimination of emotions. It is important to note that this model has been useful in feature extraction for speech recognition under noise conditions (Martínez et al 2012, 2015). This bio-inspired set along with other spectral characteristics proposed by us, are innovatively used for robust emotion recognition under several noise conditions Rao and Koolagudi (2013); Eyben et al (2013). In addition, prosodic information is incorporated in the feature vector. The results demonstrate that the proposed features significantly improve the robustness of an emotion recognition system.

In the next section the proposed methods for feature extraction and classification are introduced. Section 3 explains the emotional speech database used in the experiments. Furthermore, explains the experiments and it deals with performance measures and discussion. Finally, conclusions and future works are presented in Section 4.

2 Proposed features and classifier

The aim of this work is to propose and evaluate new sets of features for emotion recognition under noise conditions. Most of these features have never been used in robust emotion recognition (Eyben et al 2013; Rao and Koolagudi 2013; Koolagudi and Rao 2012). In addition, some parametrisations are introduced as reference baselines.

2.1 Mean of log-spectrum

We present the Mean of Log-Spectrum (MLS) coefficients for comparison purposes. It is interesting to remark that these features are evaluated under noise conditions, as an additional novelty. These are defined using the signal spectrogram as follows

$$S(k) = \frac{1}{N}\sum_{n=1}^{N} \log |v(n,k)|, \tag{1}$$

where k is a frequency band, N is the number of frames in the utterance and $v(n,k)$ is the discrete Fourier transform of the signal in the frame n. These were computed using spectrograms calculated using Hamming windows of 25 ms with a 10 ms frame shift. The first 30 MLS coefficients, corresponding to lower frequencies ($0-1200$ Hz), were considered. A previous work showed that the most useful information was found in this frequency interval (Albornoz et al 2011).

2.2 Mean of the log-auditory spectrum

In this work, we propose a set of features based on an auditory spectrogram for emotion recognition, never used for this task. Additionally, this set is evaluated under noise conditions. Yang et al (1992) proposed a model based on neurophysiological investigations at various stages of the auditory system. This model consists of two stages, the first allows to obtain an early auditory spectrogram (activity at the level of auditory nerve fibres) of the temporal signal while, the second mimics a model of primary auditory cortex in mammalian to process the spectrogram.

The first part of the model is composed of a bank of 128 cochlear filters that process the signal and obtains the 128 coefficients representing the range $[0-4000]$ Hz. This analysis stage is implemented by a bank of 128 overlapping constant-Q (QERB = 5.88) bandpass filters with centre frequencies (CF) that are uniformly distributed along a logarithmic frequency axis (x), over 5.3 octaves (24 filters/octave). The CF of the filter at location x on the logarithmic frequency axis (in octaves) is defined as

$$f_x = f_0 2^x (\mathrm{Hz}) \tag{2}$$

where f_0 is a reference frequency of 1 kHz. This quantity and frequency distribution of the filters proved to be satisfactory for the discrimination of important acoustic clues and for an appropriate reconstruction of speech signals (Chi et al 2005). As can be seen, these filters are not equally distributed in frequencies. Thus, for example, the first 71 coefficients correspond to the $[0-1220]$ Hz interval. Given that the most useful information was found in this frequency interval for the present task (Albornoz et al 2011), and in order to do a fair comparison against the MLS coefficients, only this range is

[1] Baseline feature set for the INTERSPEECH 2013 Computational Paralinguistics Evaluation Challenge.

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Feature extraction based on bio-inspired model for robust emotion recognition"
Soft Computing, Vol. 21, No. 17, pp. 5145-5158, Sep. 2017.

considered in this work. After applying this filter, the outputs are transduced into auditory-nerve patterns using a high-pass filter that represents the fluid-cilia coupling, a sigmoid function of the channel activations that represents the non-linear compression in the ionic channels, and a low-pass filter that represents hair-cell membrane leakage. Finally, the lateral inhibitory network is approximated by a first-order derivative in respect to the tonotopic (frequency) axis, which is then half-wave rectified. The output at each frequency band is then obtained by integrating this signal over a short window (Mesgarani and Shamma 2007).

We propose a new set of features using the information provided by the first stage, then the mean of the log-spectrum using the auditory spectrogram (MLSa)

$$S_a(k) = \frac{1}{N}\sum_{n=1}^{N} \log |a(n,k)|, \tag{3}$$

where k is a frequency band, N is the number of frames in the utterance and $a(n,k)$ is the k-th coefficient obtained by applying the auditory filter bank to the signal in the frame n. The MLSa were computed using auditory spectrograms calculated for windows of 25 ms without overlapping.

2.3 Principal component analysis

The principal component analysis (PCA) is a widely-known technique for dimensionality reduction (Bishop 2006) and it was used in the context of emotion recognition (El Ayadi et al 2011). This analysis allow to obtain vectors linearly uncorrelated, using an orthogonal transformation. These orthogonal vectors, called principal components, are oriented subsequently in the directions where the variance is greater. That is, the 1st component is oriented in the direction where the data present the greatest variance, the 2nd component is oriented to describe the largest possible variance perpendicular to the preceding, an so on. In this work, PCA is used to reduce the dimensionality of the data extracted from MLS and MLSa. The aim is to evaluate their usefulness in the current schema, keeping the performance of the raw features with a low-dimensionality and under noise conditions.

2.4 Smoothed MLS and MLSa

Visual inspection of MLS curves reveals similar morphologies in samples corresponding to the same class. In Figure 1 can be seen some examples of MLS curves for two emotional classes: Sadness and Disgust. In order to take advantage of these similarities, we propose to approximate the MLS and MLSa curves with polynomials and to use their coefficients as features. Firstly, a moving average filter was used in order to obtain smoother curves. After that, MLS values and frequencies of each curve were normalised between 0 and 1, individually. Finally, a polynomial of degree 7 that fits the curve in a least squares sense was computed and 8 polynomial coefficients were obtained. This information and 4 statistical characteristics of the raw curve (minimum, maximum, mean and standard deviation) were used as a feature vector. It is important to notice that the polynomials fit to the MLS and MLSa curves, therefore they are not explicitly related with temporal changes or utterances duration.

2.5 Reference features

In order to compare the proposed features with well-known ones, the MFCC and a more robust variation of the MFCC were calculated. Results obtained using theses features are showed in all the tables to discuss and compare the performances. For every emotional utterance, the first 12 MFCC were extracted using the *Hidden Markov Models Toolkit* (Young et al 2001). These features are included because they are widely-used in emotion recognition (Batliner et al 2011; El Ayadi et al 2011; Koolagudi and Rao 2012; Zeng et al 2009) and represent a baseline to evaluate the proposed representations. The MFCC were calculated using Hamming windows of 25 ms with a 10 ms frame shift. For the robust MFCC (MFCC-R), the last cosine transform did not applied and the coefficients are taken after the mel mapping. Finally, the mean of each MFCC and MFCC-R coefficient was computed along the utterance.

2.6 Incorporation of prosodic features

The use of prosodic features in emotion recognition has already been studied and discussed extensively (Adell Mercado et al 2005; Borchert and Dusterhoft 2005; Schuller et al 2004). As in almost all works, here the classic methods to calculate *Energy* and pitch (F_0) along the sentence were used (Deller Jr. et al 1993). Although many parameters can be computed from prosodic features; in this work, the mean and standard deviation of pitch and energy over the whole utterances were used, because they were the most useful in previous work (Albornoz et al 2011).

The different features were arranged in vectors with and without prosodic information. In order to present

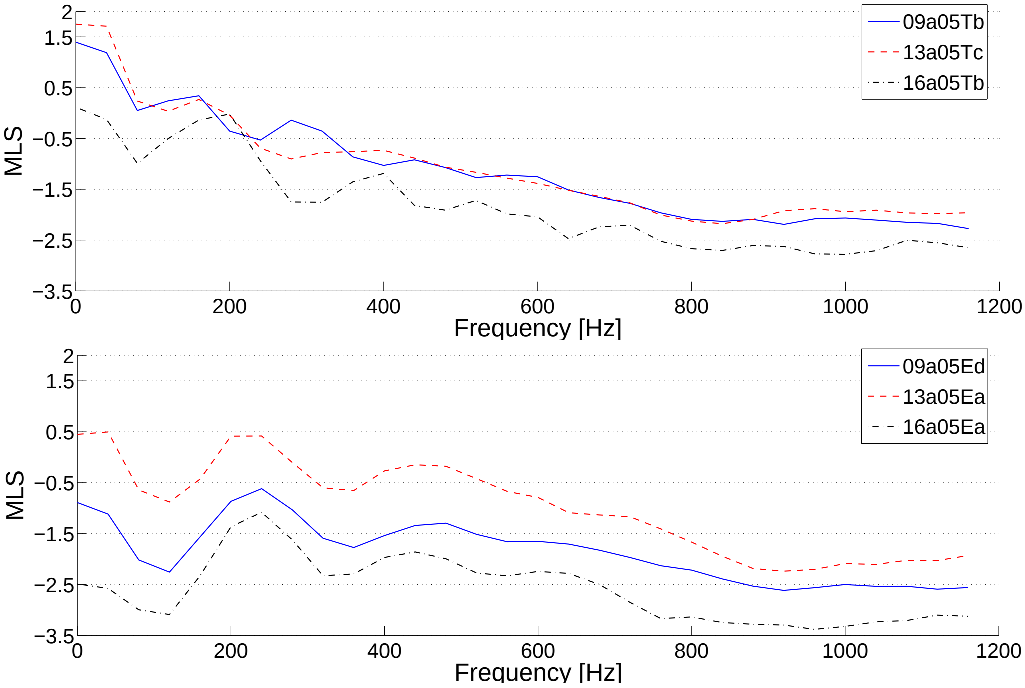

Fig. 1: MLS curves for the same sentence for 3 different speakers. Top: Sadness. Below: Disgust.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Feature extraction based on bio-inspired model for robust emotion recognition"
Soft Computing, Vol. 21, No. 17, pp. 5145-5158, Sep, 2017.

the experiments and results in a more readable form, the following notation is introduced:

- *12MFCC*: 12 mean MFCC coefficients;
- *12MFCC-R*: 12 mean MFCC filter coefficients;
- *30MLS*: 30 MLS coefficients;
- *12PCA-MLS*: 12 PCA coefficients extracted from the 30 MLS coefficients;
- *71MLSa*: 71 MLSa coefficients;
- *12PCA-MLSa*: 12 PCA coefficients extracted from the 71 MLSa coefficients;
- *12pMLS*: 12 Polynomial coefficients from 30 MLS aproximation;
- *12pMLSa*: 12 Polynomial coefficients from 71 MLSa aproximation;

The aim of selecting 12 PCA coefficients (that contain more than 95% of variance for all cases) was to keep the same dimensionality that MFCC coefficients. The features with prosodic values are pointed out using "$+P$".

2.7 Neural classifier

In order to analyse the performance of the proposed characteristics, classifiers based on neural networks were used for all the experiments. As the aim of this work is to evaluate the robustness of features for emotion recognition, we decided to use a standard MLP as classifier, with a common configuration. Classifiers based on MLP were widely used in emotion recognition (El Ayadi et al 2011; Koolagudi and Rao 2012; Schuller et al 2004). In the MLP, the nodes are fully connected between layers without connections between units in the same layer (Haykin 1998). 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. 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).

For MLP classification experiments, *Stuttgart Neural Network Simulator* (Zell et al 1998) was used. An exploration to reach the best configuration for every feature vector was performed. The initial network had one hidden layer and the exploration (changing the number of neurons N_H in the hidden layer) looked for the best number of neurons in the hidden layer N_H^*. The training of a network was stopped when it reached the generalization point with test data (Haykin 1998), and so, its test rate and trained network were kept. When all configurations for all partitions were explored, we computed an average test rate for each structure over all the partitions. The trained network that achieved the best test results was used to evaluated clean and noisy validation data.

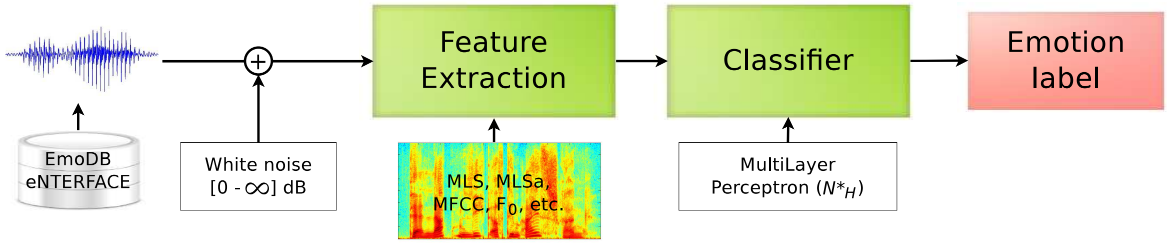

Fig. 2: Conceptual flowchart of the general whole process for the experiments.

3 Experiments

In this section, the robustness of the proposed features is evaluated. A general scheme about the whole process for the experiments is showed in Figure 2. These are divided into two groups: experiments with clean signals, that are firstly presented, and the experiments with noisy signals. For both groups of experiments, all the coefficients in vectors were normalised. Firstly, the maximum and the minimum values (for each dimension) from the training set were extracted. Then, these values were used to normalise the training vectors, clean validation vectors and all sets of noisy validation vectors. Due to the diverse sources of the features and the capabilities of the MLP, a normalisation process is really needed. We assume that the training data are sufficiently representative to train the model (as usual) and to estimate the maximum and minimum values for normalisation. In respect to noisy experiments, the same values are proposed because white noise do not generate extraneous peaks and moreover, the proposed features are averaged over the whole sentence. The MLP classifiers were evaluated considering signals artificially contaminated with additive white noise[2]. Firstly, noisy signals were generated using several signal-to-noise ratios. After that, the features were extracted as it was mentioned previously.

It is important to remark that a cross-validation method (Michie et al 1994) was implemented in all the experiments. This avoid the biased estimates of recognition error that are usually present in experiments with only one training and one test partition.

In the following, the databases used for the experiments are presented. Then, all the experiments are introduced and discussed.

3.1 Emotional Speech Corpus

In this work we used a well-known acted database *EmoDB* and a more realistic database called *eNTERFACE*. In both, all utterances belonging to the same class are labelled with the name of the class and their transcriptions are ignored. Each utterance is represented by an unique pattern in a data partition.

EmoDB

This database was developed at the Communication Science Institute of Berlin Technical University (Burkhardt et al 2005). The emotional speech signals were taken from this database because it is freely accessible[3] and it was used in several studies (Albornoz et al 2011; Borchert and Dusterhoft 2005; El Ayadi et al 2007; Schuller et al 2008; Yang and Lugger 2010). The corpus, consisting of 535 utterances expressed in German, includes sentences performed under six plain emotions, and sentences in neutral emotional state. EmoDB covers the big six emotions set except for boredom instead of surprise. The distribution, grouped by emotion class, is:

– Anger: 127 utterances;
– Boredom: 81 utterances;
– Disgust: 46 utterances;
– Fear: 69 utterances;
– Joy: 71 utterances;
– Sadness: 62 utterances;
– Neutral: 79 utterances.

The unbalanced distribution of classes is important (24% of the set is represented by *anger* class). Methods that use this kind of data usually obtain biased results. Almost all works did not address this issue and their results are hardly comparable. This is a really important topic in most types of classifiers.

In the same way as in (Albornoz et al 2011), the dataset was balanced by equalizing the size of the classes and ten partitions were generated. The same number of samples for all classes in each partition were

[2] Using Matlab.

[3] http://pascal.kgw.tu-berlin.de/emodb/.

randomly selected (46 x 7 = 322 utterances) without speaker considerations. In the classification experiments, the data partitions were randomly divided in 80% for *training* and the remaining 20% was left for *final validation*. The training data in turn was randomly separated in 60% for training and 20% for the generalization test.[4] This validation methodology includes a cross-validation using the ten partitions. This is the same schema defined in (Albornoz et al 2011) and thus, the results are directly comparable.

The previous validation methodology did not ensure speaker independence (in the sense of leave-k-speaker out) and we attempted to deal with this issue. It is important to mention that the distribution of the utterances in the database (for different speakers and types of emotions) is very irregular (Burkhardt et al 2005). Therefore, a leave-one-out scheme is impossible because sometimes there is not enough utterances for validation. However, none previous work - to our best knowledge - has taken into account this important issue. We decided to propose a more exigent scheme using leave-two-out (one female + one male), which allows the balance. In order to do that, we generated 5 partitions combining the couples in the more balanced way.

eNTERFACE

The eNTERFACE database has 1260 utterances and it is freely accessible[5] (Martin et al 2006). This is an audio-visual emotion database that contains 42 subjects whom express 5 sentences for each of 6 archetypal emotions: happiness, sadness, surprise, anger, disgust and fear. Utterances are expressed in English as reaction to a particular situation by people coming from 14 different countries. In this work, the audio signals were extracted in *wav* format (using ffmpeg 1.0.6.) from the *avi* format.

For speaker independent experiments, we defined 7 partitions that contain: Training (900 sentences - 30 speakers), Test (180 sentences - 6 speakers) and Final Validation (180 sentences - 6 speakers).

3.2 Experiments without noise

In order to set up a reference system, the proposed feature vectors were evaluated without noise. Table 1 shows the classification accuracy for the proposed feature vectors. To estimate if prosody is useful in complementing these features, we evaluate the extension of feature vectors with prosodic information (Table 2). Several configurations for the MLP were tested and then, the best ones were reported in the tables. The *12pMLS* and *12pMLSa* have obtained the poorer performance, even adding prosodic information. For *12MFCC* and *12MFCC-R*, the performances were suitable showing significant improvements when these were used with prosody. Vectors based on MLS obtained satisfactory performances that exhibit enhancements when prosodic information was added. It is interesting to note that *30MLS* and *12PCA-MLS* achieved similar results (with and without prosody). It would demonstrate that PCA is useful to perform a dimensional reduction of MLS. The *71MLSa* showed an unsatisfactory behaviour, and the 12 PCA coefficients based on MLSa obtained a similar result. However, when the 71 MLSa coefficients were used with prosodic information (*71MLSa + P*), the classification rate was improved more than 8 % (absolute). Furthermore, it can be observed that the *12PCA-MLSa + P* have reached the best classification rate, improving almost 18 % (absolute) the classification rate achieved without prosody (*12PCA-MLSa*). This result indicates once again that PCA is suitable to reduce the dimensionality of spectral coefficients (MLS and MLSa) due to it retain the discriminative information, whereas it points out that prosodic and MLSa information should be used together. In addition, PCA

Table 1: Classification rate for clean signals without prosody. Accuracy (Acc) is computed with validation data.

Feature Vector	Acc [%]
12MFCC	63.17
12MFCC-R	58,41
30MLS	56.83
12PCA-MLS	56.51
71MLSa	48.89
12PCA-MLSa	48.73
12pMLS	48.41
12pMLSa	44.61

Table 2: Classification rate for clean signals with prosody. Accuracy (Acc) is computed with validation data.

Feature Vector	Acc [%]
12MFCC + P	65.71
12MFCC-R + P	64,92
30MLS + P	59.37
12PCA-MLS + P	59.21
71MLSa + P	56.35
12PCA-MLSa + P	**66.19**
12pMLS + P	52.70
12pMLSa + P	49.37

[4] Each partition has 196 utterances for training, 63 utterances for generalization test and 63 utterances for the final validation.

[5] Available at http://www.enterface.net/enterface05.

sinc(i) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Feature extraction based on bio-inspired model for robust emotion recognition"
Soft Computing, Vol. 21, No. 17, pp. 5145-5158, Sep. 2017.

Table 3: Performance of features under noise conditions (EMODB: 10-fold cross validation). Accuracy in [%].

SNR	*12MFCC*	*12MFCC-R*	*30MLS*	*12PCA-MLS*	*71MLSa*	*12PCA-MLSa*
∞ dB	**63.17**	58.41	56.83	56.51	48.89	48.73
40 dB	34.29	52.06	**57.94**	56.03	49.05	49.68
35 dB	30.16	44.28	**57.46**	54.60	47.30	49.52
30 dB	28.41	35.40	53.02	**53.81**	45.08	47.94
25 dB	24.60	25.55	47.30	**50.79**	41.59	45.40
20 dB	23.02	19.68	36.19	**44.60**	34.92	41.75
15 dB	20.00	18.57	30.32	33.49	28.09	**34.13**
10 dB	17.62	16.99	23.65	26.19	23.81	**29.84**
5 dB	16.51	16.35	18.89	20.48	20.79	**26.03**
0 dB	16.03	16.19	16.03	16.51	16.99	**20.79**

Table 4: Robust classification for features with prosody (EMODB: 10-fold cross validation). Accuracy in [%].

SNR	*12MFCC+P*	*12MFCC-R+P*	*30MLS+P*	*12PCA-MLS+P*	*71MLSa+P*	*12PCA-MLSa+P*
∞ dB	65.71	64.92	59.37	59.21	56.35	**66.19**
40 dB	39.68	53.97	**58.89**	56.67	53.97	42.06
35 dB	36.35	46.51	**58.89**	57.14	51.90	42.06
30 dB	33.49	34.44	52.07	**53.81**	50.32	42.22
25 dB	30.00	25.40	46.67	**49.05**	46.98	40.64
20 dB	26.35	21.75	40.00	**43.97**	41.43	38.41
15 dB	25.08	19.05	30.63	**34.13**	33.18	32.06
10 dB	22.54	17.46	23.18	**26.51**	**26.51**	26.35
5 dB	20.00	15.72	20.16	23.33	22.07	**24.60**
0 dB	18.57	15.24	18.41	19.84	20.48	**22.70**

allowed to have (at the most) the same dimensionality of MFCC. It is an interesting and promising result because these features are new for the emotion recognition task.

3.3 Experiments with noise

In the experiments with noisy signals, the features that showed a poor performance in preliminary experiments without noise were not included. The first experiments were performed using the EMODB and 10-fold cross validation. The classifiers trained with clean signals were tested with noisy signals and the performance is presented in Table 3.

In the first column, signal-to-noise ratios are displayed and the remaining columns show the performance of every feature vector. As can be observed in this table, similar to previous works, *12MFCC* works fine without noise but its behaviour degrades abruptly in noise presence. The *12MFCC-R* features are more robust than *12MFCC* for low levels of noise. The set of proposed features obtained a significant improvement in classification rates. It should be noted, for example, an absolute increment of more than 23 % (from 34.29 to 57.94) with 40 dB and an absolute increment of more than 27 % (from 30.16 to 57.46) with 35 dB (with regard to MFCC). After that dB range, the PCA becomes beneficial to reduce noise effects, whereas that allows dimensional reduction. Thus, the absolute improvement is greater than 21% (from 23.02 to 44.60, from 24.60 to 50.79 and from 28.41 to 53.81) in $20-30$ dB range. Finally, a relative classification enhancement about $60-70$ % (from 16.51 to 26.03, from 17.62 to 29.84, and from 20.00 to 34.13) is obtained for low signal-to-noise ratio ([$5-15$] dB). Meanwhile the relative improvement is about 30 % (from 16.03 to 20.79) for 0 dB. The parametrisation based on the bio-inspired model shows a robust behaviour against high presence of noise. Unlike the MFCC, the proposed features are affected in conditions of high SNR (up to 25 dB). The significant improvements, that can be observed when the dimensionality of features is reduced using PCA, would suggest that these representations allow to keep in a separate manner the signal information and the noise. Furthermore, the non-linear filter bank of the auditory model can obtain relevant information even in the presence of higher levels of noise.

Using the EMODB and 10-fold cross validation, the feature sets with prosodic information were evaluated and the results are presented in Table 4. Results show that prosodic information improves the classification for all representations, using clean signals (more than 17% in *12PCA-MLSa* case). This information was not always useful under noisy conditions and it could be attributable to the non-robust method used for prosodic computation. The representations show a similar behaviour in Table 3 and 4. In this way, the PCA coeffi-

Table 5: Robust classification for features with prosody (EMODB). Classifiers trained with 0 dB SNR). Accuracy in [%].

SNR	*12MFCC+P*	*12MFCC-R+P*	*30MLS+P*	*12PCA-MLS+P*	*71MLSa+P*	*12PCA-MLSa+P*
0 dB	**56.67**	51.11	48.89	52.54	47.94	42.70
5 dB	44.13	32.38	44.29	**47.46**	41.27	41.27
10 dB	35.56	27.78	39.21	**39.37**	33.97	**39.37**
15 dB	28.73	23.97	34.45	34.29	28.57	**34.76**
20 dB	25.56	22.54	30.95	31.75	28.09	**32.54**
25 dB	24.13	21.90	29.52	29.68	26.67	**31.27**
30 dB	22.38	20.64	28.57	28.57	25.56	**29.68**
35 dB	22.54	19.52	27.14	**28.73**	23.97	28.09
40 dB	21.75	19.21	26.51	27.14	23.49	**27.78**

Table 6: Robust classification for features without prosody (EMODB: leave-2-out scheme). Accuracy in [%].

SNR	*12MFCC*	*12MFCC-R*	*30MLS*	*12PCA-MLS*	*71MLSa*	*12PCA-MLSa*
∞	**55.07**	46.22	54.23	53.97	46.35	47.99
40 dB	34.15	45.50	53.59	**55.14**	45.65	48.34
35 dB	31.87	42.36	52.03	**54.26**	44.90	47.58
30 dB	28.85	36.38	52.04	**52.59**	44.57	48.94
25 dB	26.68	27.30	50.11	**51.68**	43.86	45.77
20 dB	23.91	24.23	45.88	**49.11**	40.44	45.59
15 dB	21.80	22.08	39.93	39.70	41.20	**41.52**
10 dB	20.35	20.39	34.70	31.66	36.22	**36.35**
5 dB	19.90	19.83	25.80	21.78	**32.01**	30.17
0 dB	18.40	19.27	15.97	19.91	**28.68**	26.56

cients obtained the best performance and furthermore, they provide low dimensional vectors. Once again, the PCA extracted from the bio-inspired parametrisation presents a robust behaviour against high presence of noise. Thus, the relative classification improvement is about 22 % for 0 dB (from 18.57 to 22.70 %).

As introduced previously, the most relevant information for emotion recognition was found between 0 and 1200 Hz. The good results can be attributable, on one hand, to the better resolution of the filter bank of the bio-inspired model in this frequency range that imitates frequency selectivity of the basilar membrane and, on the other hand, to the filters and approximations that simulate the remainder behaviour of the early auditory system. Moreover, the dimensional reduction using PCA keeps relevant characteristics of emotional signals disregarding the noise information while it reduces the dimensionality of the input vector for the classifier.

The previous experimental results were obtained with classifiers trained with clean signals and validated with contaminated signals for several SNR. A hard experiment was performed using classifiers trained using noisy signals (0 dB SNR) and validated with noisy signals for diverse SNR. The best MLP structures obtained for clean signals, for each feature set, were considered. The classifiers were validated using EMODB and ten-fold cross validation. In Table 5 the results are presented. For this scheme, the behaviour of MLS and MLSa features are relatively good and when the PCA was used, the best performances were obtained. The MFCC features obtained the best performance when the classifier is trained and validated using signals with the same SNR.

Other set of experiments to deal with speaker independence was performed using the EMODB. In order to carry out the experiments, a leave-k- speakers out scheme was considered as was mentioned in Section 3.1. In the same way that previous experiments the classifiers were trained with clean signals and then were validated for several SNR using 5-fold cross validation. Tables 6 and 7 present the results. As can be seen, the performance decreases for ∞ dB configuration respect to previous results. The features based on MLS have the best performance for high SNR, while the MLSa related features have the best performance for low SNR. Additional information is showed graphically in Figures 3 and 4. In both can be observed the relative improvement of classification rate with regard to *12MFCC* and *12MFCC + P*, respectively. It is noticeable the significant relative improvement obtained in almost cases. For example, as can be seen in Figure 3 the proposed features improved the *12MFCC* more than 30 % up to 10 dB, reaching a maximum over 100 % of relative improvement. Although all features experiment a decay to 0 dB, MLSa obtains 55.90 % (relative) more than MFCC (from 18.40 to 28.68) with 0 dB SNR. The *12MFCC-R* shows again a more robust behaviour

Table 7: Robust classification for features with prosody (EMODB: leave-2-out scheme). Accuracy in [%].

SNR	*12MFCC+P*	*12MFCC-R+P*	*30MLS+P*	*12PCA-MLS+P*	*71MLSa+P*	*12PCA-MLSa+P*
∞	54.56	43.76	**56.07**	52.17	52.38	50.56
40 dB	35.71	44.18	**53.85**	52.64	51.09	47.41
35 dB	34.34	42.19	**53.22**	52.17	48.84	46.27
30 dB	32.79	35.94	**51.53**	51.36	49.01	44.98
25 dB	31.87	29.28	47.53	**48.01**	44.95	43.58
20 dB	31.91	26.21	43.81	**45.30**	42.33	41.45
15 dB	32.92	25.25	39.09	36.90	**41.25**	39.98
10 dB	32.30	20.23	32.89	30.27	36.53	**37.34**
5 dB	29.62	16.24	28.79	28.71	34.73	**35.83**
0 dB	28.33	16.35	22.41	21.45	**34.34**	33.85

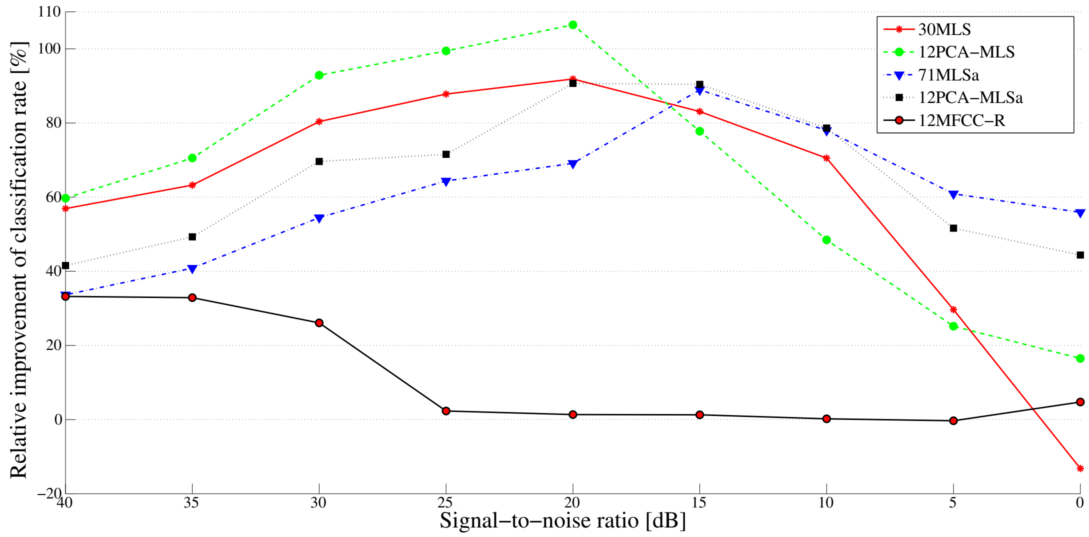

Fig. 3: Robust classification for features without prosody (EMODB: leave-2-out scheme). Classification rates are relative to *12MFCC*.

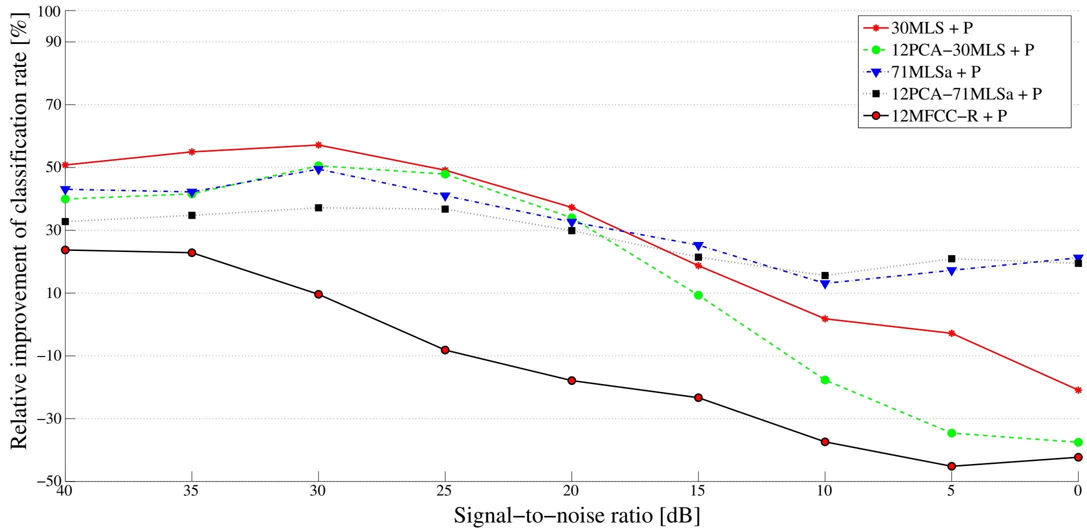

Fig. 4: Robust classification for features with prosody (EMODB: leave-2-out scheme). Classification rates are relative to *12MFCC+P*.

sinc(*i*) Research Institute for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Feature extraction based on bio-inspired model for robust emotion recognition"
Soft Computing, Vol. 21, No. 17, pp. 5145-5158, Sep. 2017.

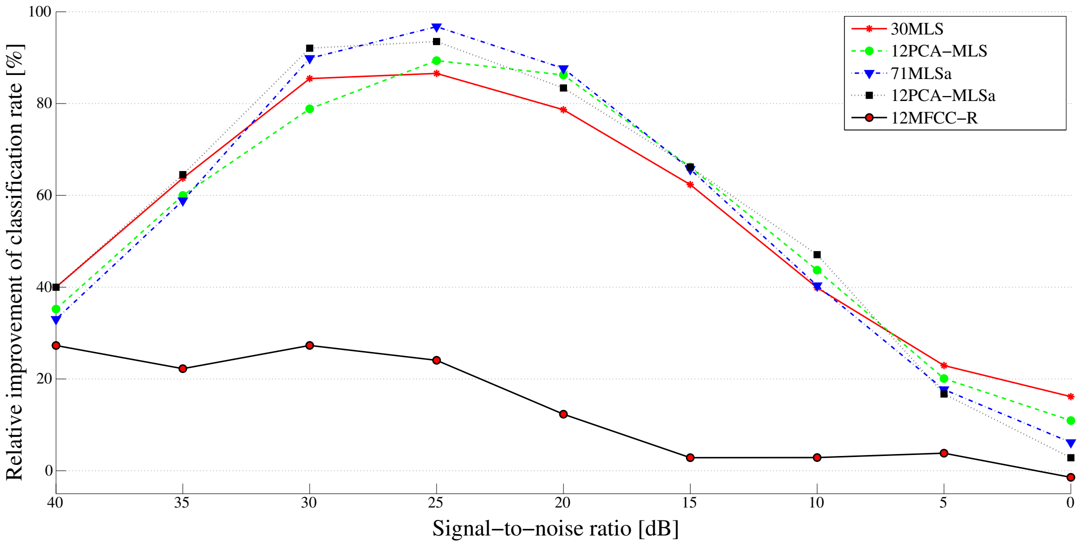

Fig. 5: Robust classification for features without prosody (eNTERFACE). Classification rates are relative to *12MFCC*.

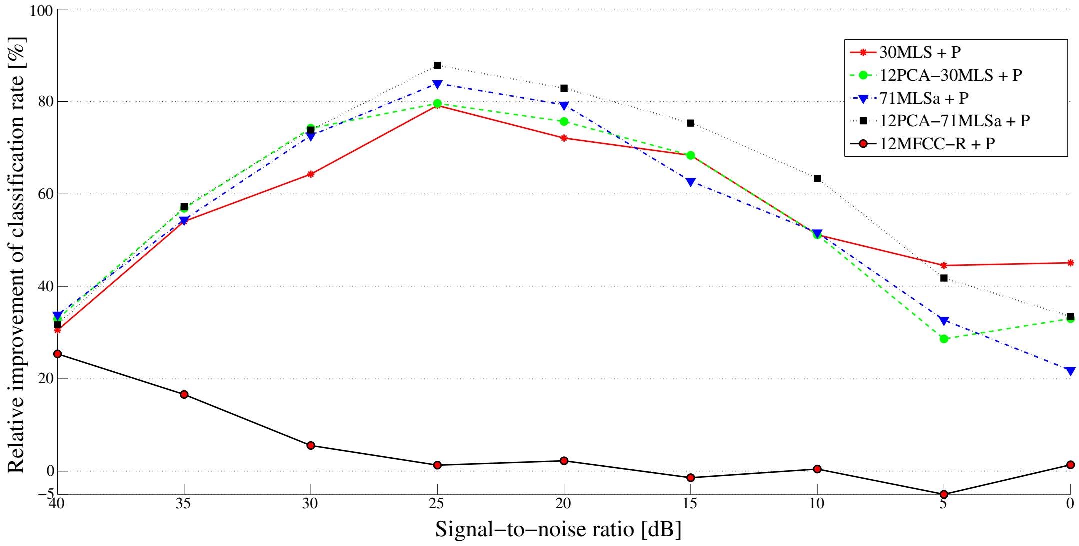

Fig. 6: Robust classification for features with prosody (eNTERFACE). Classification rates are relative to *12MFCC + P*.

respect to MFCC for high SNR. Figure 4 presents the results of feature vectors with prosody. In this case, the incorporation of prosody is more beneficial to the reference features (*12MFCC + P*). While the gap of relative improvement decreases respect to the previous figure, the proposed features show a superior performance (excepting MLS features in 0−10 dB range where the white noise produces a great masking on the traditional spectrum). In both figures can be observed that proposed parametrisations have evidenced a good performance under noisy conditions. Features based on MLS have a good behaviour to high SNR, whereas the bio-inspired features reach the best performance under conditions with higher level of noise. These experiments confirm that the results are promising also in a speaker independence scheme.

In order to obtain a more realistic impression of the previous results, we considered a more spontaneous database in a different language (eNTERFACE). For these experiments, a 7-fold cross validation scheme was implemented. Using the clean signals, different MLP structures were explored in order to find the best network configuration for each set of extracted features. The classifiers, trained with clean signals, were validated with contaminated signals for several SNR. In Figures 5 and 6 can be observed the relative improvement of classification rate with regard to *12MFCC* and *12MFCC + P*, respectively. In these experiments, the proposed parametrisations present a significant improvement in all cases under noisy conditions. For example, as can be seen in Figure 5 the proposed features improved the *12MFCC* more than 30 % in $40-10$ dB range, reaching maximums over 80 % of relative improvement for 30, 25 and 20 dB. Figure 6 shows that the incorporation of prosody is beneficial to the reference features (*12MFCC + P*) in $40-20$ dB range, while the proposed features always show a superior performance. Figure 6 shows that, although the incorporation of prosody is a bit more beneficial to the reference features (*12MFCC + P*) in $40-20$ dB range, the proposed features always show a superior performance. Moreover, the gap of relative improvement increases in $15-0$ dB range, outperforming 45 % relative respect the to MFCC with 0 dB SNR (from 17.07 to 24.76). On the other hand, as in previous experiments the *12MFCC-R* showed a more robust performance respect to MFCC for low levels of noise. Both figures show that the proposed parametrisations have a similar good performance. As can be seen, the bio-inspired features also obtained promising results using a more realistic corpus in a different language.

4 Conclusions and future work

In this paper, we present a novel set of features based on a bio-inspired model for emotion recognition. These features are obtained from a time-frequency analysis computed with an auditory model. Also, PCA was used to get lower dimensional vectors from the proposed features. These features were evaluated under noise conditions, which is a current challenge for this recognition task.

Results showed that the MLS and MLSa features, and these combined with prosody, improve the accuracy under these noise conditions. Moreover, the PCA proved to be useful to extract the most relevant information from MLS and MLSa because it allowed to improve the results with low dimensional data. As well as in previous works, the MFCC showed a good performance with clean signals but it degrades a lot in noise presence. The *MFCC-R* showed to be a reliable alternative to MFCC for low levels of noise. On the other hand, the features proposed in this work always improved the classification under noise conditions. The results pointed out that the features based on spectral information have a very good performance for low amount of noise, whereas the characteristics computed using the auditory model have shown to be more robust for low signal-to-noise ratios. The results were validated in a speaker independent scheme and with two emotional speech corpora.

In future work the proposed parametrisation could be used to improve the performance of hierarchical classifiers with noisy signals. Furthermore, we will explore the performance of classifiers with another type of noises (pink, bubble and others non-stationary) and types of noise interaction models (like the convolutive in reverberant rooms).

Acknowledgements The authors wish to thank to *Agencia Nacional de Promoción Científica y Tecnológica* and *Universidad Nacional de Litoral* (with PAE 37122, PACT 2011 #58, CAI+D 2011 #58-511) and *Consejo Nacional de Investigaciones Científicas y Técnicas* (CONICET) from Argentina, for their support.

Compliance with Ethical Standards

Conflict of Interest: The authors declare that they have no conflict of interest.
Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

Adell Mercado J, Bonafonte Cávez A, Escudero Mancebo D (2005) Analysis of prosodic features: towards modelling of emotional and pragmatic attributes of speech. Procesamiento de Lenguaje Natural (35):277–283

Albornoz EM, Milone DH (2016) Emotion recognition in never-seen languages using a novel ensemble method with emotion profiles. IEEE Transactions on Affective Computing PrePrints(0):–, DOI 10.1109/TAFFC.2015.2503757

Albornoz EM, Milone DH, Rufiner HL (2011) Spoken emotion recognition using hierarchical classifiers. Computer Speech & Language 25(3):556–570

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 CM (2006) Pattern Recognition and Machine Learning, 1st edn. 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 (NLP-KE), pp 147–151

Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B (2005) A Database of German Emotional Speech. In: Proceedings of 9th European Conference on Speech Communication and Technology (Interspeech), pp 1517–1520

Chanel G, Kierkels JJ, Soleymani M, Pun T (2009) Short-term emotion assessment in a recall paradigm. International Journal of Human-Computer Studies 67(8):607 – 627

Chi T, Ru P, Shamma SA (2005) Multiresolution spectrotemporal analysis of complex sounds. Journal of the Acoustical Society of America 118(2):887–906

Chin YH, Lin SH, Lin CH, Siahaan E, Frisky A, Wang JC (2014) Emotion Profile-Based Music Recommendation. Proc 7th International Conference on Ubi-Media Computing and Workshops (UMEDIA) pp 111–114

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

Cummins N, Scherer S, Krajewski J, Schnieder S, Epps J, Quatieri TF (2015) A review of depression and suicide risk assessment using speech analysis. Speech Communication 71:10–49

Dellaert F, Polzin T, Waibel A (1996) Recognizing Emotions in Speech. In: Proceedings of International Conference on Spoken Language Processing (ICSLP), Philadelphia, PA, vol 3, pp 1970–1973

Deller Jr JR, Proakis JG, Hansen JH (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, Lecture Notes in Computer Science, vol 4441/2007, Springer-Verlag, Berlin, Heidelberg, chap 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: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol 4, pp IV–957–IV–960

El Ayadi M, Kamel M, Karray F (2011) Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition 44(3):572–587

Eyben F, Schuller B, Rigoll G (2012) Improving generalisation and robustness of acoustic affect recognition. In: Proceedings of the 14th ACM international conference on Multimodal interaction, ACM, New York, NY, USA, ICMI '12, pp 517–522

Eyben F, Weninger F, Schuller B (2013) Affect recognition in real-life acoustic conditions-a new perspective on feature selection. In: In Proceedings of Interspeech, pp 2044–2048

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

Giakoumis D, Tzovaras D, Hassapis G (2013) Subject-dependent biosignal features for increased accuracy in psychological stress detection. International Journal of Human-Computer Studies 71(4):425 – 439

Han Z, Lun S, Wang J (2012) A study on speech emotion recognition based on CCBC and neural network. In: Proceedings of International Conference on Computer Science and Electronics Engineering (ICCSEE), IEEE Computer Society, vol 2, pp 144–147

Haykin S (1998) Neural Networks: A Comprehensive Foundation, 2nd edn. Prentice Hall

Iliev AI, Scordilis MS, Papa JP, Falcão AX (2010) Spoken emotion recognition through optimum-path forest classification using glottal features. Computer Speech & Language 24(3):445–460

Kandali A, Routray A, Basu T (2010) Vocal emotion recognition in five languages of Assam using features based on MFCCs and Eigen Values of Autocorrelation Matrix in presence of babble noise. In: Proceedings of National Conference on Communications (NCC), pp 1–5

Kapoor A, Burleson W, Picard RW (2007) Automatic prediction of frustration. International Journal of Human-Computer Studies 65(8):724 – 736

Kim J (2007) Robust Speech Recognition and Understanding, I-Tech Education and Publishing, Vienna, Austria, chap 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

Koolagudi S, Rao K (2012) Emotion recognition from speech using source, system, and prosodic features. International Journal of Speech Technology 15:265–289

Lazarus R (2001) Appraisal Processes in Emotion: Theory, Methods, Research (Series in Affective Science), Oxford University Press, USA, chap 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 YL, Wei G (2005) Speech emotion recognition based on HMM and SVM. In: Proceedings of International Conference on Machine Learning and Cybernetics, 2005., vol 8, pp 4898–4901

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

Lugger M, Yang B (2008) Speech Recognition, InTech, chap Psychological Motivated Multi-Stage Emotion Classification Exploiting Voice Quality Features, Speech Recognition

Martin O, Kotsia I, Macq B, Pitas I (2006) The enterface'05 audio-visual emotion database. In: Proceedings of the 22nd International Conference on Data Engineering Workshops, IEEE Computer Society, Washington, DC, USA, ICDEW '06, pp 1517–1520

Martínez C, Goddard J, Milone D, Rufiner H (2012) Bioinspired sparse spectro-temporal representation of speech for robust classification. Computer Speech & Language 26(5):336–348

Martínez C, Goddard J, Di Persia L, Milone D, Rufiner H (2015) Denoising sound signals in a bioinspired non-negative spectro-temporal domain. Digital Signal Processing: A Review Journal 38:22–31

Mesgarani N, Shamma S (2007) Denoising in the domain of spectrotemporal modulations. EURASIP J Audio Speech Music Processing 2007(3):1–8

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

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

Noguerias A, Moreno A, Bonafonte A, Mariño J (2001) Speech Emotion Recognition Using Hidden Markov Models. In: Proceedings of European Conference on Speech Communication and Technology (Eurospeech), Aalborg, Denmark, pp 2679–2682

Pao TL, Liao WY, Chen YT, Yeh JH, Cheng YM, Chien C (2007) Comparison of Several Classifiers for Emotion Recognition from Noisy Mandarin Speech. In: Proceedings of 3rd International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIHMSP), vol 1, pp 23–26

Rao KS, Koolagudi SG (2013) Robust Emotion Recognition using Spectral and Prosodic Features. Springer Science & Business Media

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), pp I–577–80

Schuller B, Seppi D, Batliner A, Maier A, Steidl S (2007) Towards more reality in the recognition of emotional speech. In: Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol 4, pp IV–941 –IV–944

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: Proceedings of IEEE International Conference on Multimedia and Expo, pp 1333–1336

Schuller B, Steidl S, Batliner A, Burkhardt F, Devillers L, Müller C, Narayanan S (2013) Paralinguistics in speech and language - state-of-the-art and the challenge. Computer Speech & Language 27(1):4–39, Special issue on Paralinguistics in Naturalistic Speech and Language

Schuller BW, Weninger F (2012) Ten recent trends in computational paralinguistics. In: Esposito A, Esposito AM, Vinciarelli A, Hoffmann R, Müller VC (eds) Cognitive Behavioural Systems, Lecture Notes in Computer Science, vol 7403, Springer Berlin Heidelberg, pp 35–49

Shamma SA, Chadwick RS, Wilbur WJ, Morrish KA, Rinzel J (1986) A biophysical model of cochlear processing: Intensity dependence of pure tone responses. The Journal of the Acoustical Society of America 80(1):133–145

Shojaeilangari S, Yau WY, Nandakumar K, Li J, Teoh EK (2015) Robust representation and recognition of facial emotions using extreme sparse learning. Image Processing, IEEE Transactions on 24(7):2140–2152

Sztahó D, Imre V, Vicsi K (2011) Automatic classification of emotions in spontaneous speech. In: Esposito A, Vinciarelli A, Vicsi K, Pelachaud C, Nijholt A (eds) Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues, Lecture Notes in Computer Science, vol 6800, Springer Berlin Heidelberg, pp 229–239

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, Tampere, Finland, pp 100–102

Tawari A, Trivedi M (2010) Speech emotion analysis in noisy real-world environment. In: Proceedings of 20th International Conference on Pattern Recognition (ICPR), pp 4605–4608

Truong KP, van Leeuwen DA (2007) Automatic discrimination between laughter and speech. Speech Communication 49(2):144–158

Wagner J, Vogt T, Andr E (2007) A Systematic Comparison of Different HMM Designs for Emotion Recognition from Acted and Spontaneous Speech. In: Paiva A, Prada R, Picard R (eds) Affective Computing and Intelligent Interaction, Lecture Notes in Computer Science, vol 4738, Springer Berlin Heidelberg, pp 114–125

Wang Y, Guan L, Venetsanopoulos AN (2012) Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. IEEE Transactions on Multimedia 14(3):597–607

Wöllmer M, Kaiser M, Eyben F, Schuller B, Rigoll G (2013) Lstm-modeling of continuous emotions in an audiovisual affect recognition framework. Image and Vision Computing 31(2):153 – 163

Xiao Z, Dellandréa E, Dou W, Chen L (2009) Recognition of emotions in speech by a hierarchical approach. In: Proceedings of 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

Yang X, Wang K, Shamma SA (1992) Auditory representations of acoustic signals. IEEE Transactions on Information Theory 38(2):824–839

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 KU, Soyez T, Schmalzl M, Sommer T, Hatzigeorgiou A, Posselt D, Schreiner T, Kett B, Clemente G (1998) SNNS (Stuttgart Neural Network Simulator). Stuttgart, Germany, SNNS User Manual Version 4.0

Zeng Z, Pantic M, Roisman G, Huang T (2009) A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions. IEEE Transactions on Pattern Analysis and Machine Intelligence 31(1):39–58