XV Reunión de Trabajo en Procesamiento de la Información y Control, 16 al 20 de septiembre de 2013

Robust Emotion Recognition using Bio-inspired Features

Enrique M. Albornoz[†,‡,*], Diego H. Milone[†,‡], Hugo L. Rufiner[†,‡]

†Centro de Investigación en Señales, Sistemas e INteligencia Computacional (SINC(i))
Facultad de Ingeniería y Ciencias Hídricas, Universidad Nacional del Litoral
‡Consejo Nacional de Investigaciones Cientíﬁcas y Técnicas (CONICET)
** emalbornoz@fich.unl.edu.ar*

***Abstract*— Several bio-inspired representations have been applied to artificial systems for speech processing. In this work, an auditory signal representation is used to obtain bio-inspired features of emotional speech signals. These features, together with other spectral and prosodic features, is used for emotion recognition under noise conditions. Multilayer perceptrons were trained as classifiers and results were compared to the well-known mel frequency cepstral coefficients with the same type of classifiers. Results show that using the proposed representations, it is possible to significantly improve the robustness of an emotion recognition system.**

***Keywords*— Emotion recognition, bio-inspired representations, multilayer perceptrons.**

1 Introduction

Emotion recognition is a multi-disciplinary research area because the emotions are motivated by social and psychological facts and these can be perceived in speech signals, in facial expressions, in body posture, in biosignal as electrocardiograph, blood pressure and electroencephalogram, among others. Real-life applications have motivated the researchers. For example in detection of fear in abnormal situations (for security applications) [1], support of semi-automatic diagnosis of psychiatric diseases [2] and detection of emotional attitudes of children in dialog interactions with computers [3].

Emotions are expressed in more than one modality and several studies have explored multimodal systems in the context of emotion recognition [4, 5]. In spite of their good results, the use of speech signals remains the most feasible option because the methods to record and use these signals are simple, not invasive and possible in most real applications. The problem of feature extraction of speech could be focussed on different aspects: speech production, characteristics of speech signals, speech perception, etc. However, most of the researchers have addressed their analysis to speech prosodic features and spectral information: Mel frequency cepstral coefficients (MFCCs), linear prediction cepstral coefficients (LPCCs), perceptual linear prediction coefficients (PLPCs), and formant measures, among others [6, 7, 8].

For classification stage, several standard techniques have been explored: Gaussian Mixture Models (GMM), Hidden Markov Models (HMM), Multilayer Perceptron (MLP), Support Vector Machines (SVM), k-nearest neighbor (k-NN), Bayesian classifiers [7, 9, 10]. At present, the combination of standard methods have become the focus of state-of-the-art studies [7, 11]. Morrison et al. [12] applied stacked generalization and unweighted vote to emotion recognition. In [13], a two-stage classifier using SVM and HMM was proposed. Schuller et al. [14] presented a multiple stage classifier using SVM. A hierarchical model and a binary multi-stage classifier guided by the dimensional emotion model were proposed in [15]. Albornoz et al. [16] developed a hierarchical classifier using clusters of emotions defined by spectral and prosodic features.

A current challenge that has not received enough attention is the robust emotion recognition [11, 17]. Some previous works have used databases recorded in real environments [18, 19] and other researchers have explored the controlled noise addition [20, 21]. Sztahó et al. [18] proposed some acoustical features and a SVM classifier for emotion classification using a database with spontaneous telephone conversations played by actors, but they did not perform an objective evaluation of noise level. In [19], the emotion recognizer is preceded by a module for adaptive noise cancellation. They evaluated the performance with white Gaussian noise and noises in a car for different scenarios. White noise addition was explored for two acted and one spontaneous databases in [21]. For each database, they extracted 4000 acoustic features, and 25 additional suprasegmental prosodic features for spontaneous database. Using random forests as classifier, all features and the best features computed with Sequential Forward Floating Selection (SFFS) were explored. In [20], prosodic and quality features with a MLP classifier were used. Authors performed the Canonical Correlation Based on Compensation, after feature extraction stage. Classification shows a good performance under 10 dB SNR, although they did not indicate the noise type.

In our work, a new set of bio-inspired features is proposed for emotion recognition which are computed using the auditory model proposed by Shamma et al. [22]. This and another set of features based on spectral analysis [16] are used for emotion recognition under several

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner, "Robust Emotion Recognition using Bio-inspired Features"
Anales de la XV Reunión de Procesamiento de la Información y Control, pp. 94-99, sep, 2013.

noise conditions. Prosodic information is also used and the classifiers are developed using a standard MLP.

In the next section the proposed methods for feature extraction and classification are presented. Section 3 explains the emotional speech database and the experiments. Furthermore, it describes the performance results and presents the discussion. Finally, conclusions and future works are presented.

2 Proposed method

In this work, we propose and evaluate different features for emotion recognition under noise conditions. These features are novel in robust emotion recognition [6, 7]. A classifier based on neural networks was used in all the experiments because our objective is to compare the features.

Mean of the log-auditory spectrum

In this work, a set of features based in the auditory spectrogram is proposed for emotion recognition. Yang et al. proposed a model based on neurophysiological investigations at various stages of the auditory system [23]. This model consists of two stages, the first allows to obtain an early auditory spectrogram of the temporal signal at the level of auditory nerve fibers. The second stage 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 cochlear filters that process the signal and obtains 128 coefficients representing the range of 0 to 4000 Hz. This analysis stage is implemented by a bank of 128 overlapping constant-Q (QERB = 5.88) bandpass filters with center frequencies (CF) that are uniformly distributed along a logarithmic frequency axis, 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

$$f_x = f_0 2^x (\text{Hz}) \tag{1}$$

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 [24]. As can be seen, these filters are not equally distributed in frequencies. Thus, for example, the first 71 coefficients correspond to the $[0 - 1200]$ Hz interval. Given that for the present task the most useful information was found in this frequency interval [16], only this range is 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. Then, it uses 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 with 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 [25].

We propose a new set of features using this information: the mean of the log-spectrum using the auditory spectrogram (MLSa), defined as

$$S_a(k) = \frac{1}{N} \sum_{n=1}^{N} \log |a(n,k)|, \tag{2}$$

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.

Mean of the log-spectrum

The mean of the log-spectrum (MLS) coefficients were used for comparison purposes and to evaluate its behaviour under noise conditions. These are defined using the signal spectrogram as follow

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

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 from 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 as in [16].

The principal component analysis (PCA) is a widely-known technique for dimensionality reduction [26] and it was used in the context of emotion recognition [7]. This analysis allows 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. In this work, PCA is used to reduce the dimensionality of the data extracted from MLS and MLSa.

Mel frequency cepstral coefficients

This parameterisation, widely-known in emotion recognition, is included as baseline [7, 6, 11]. For every emotional utterance, the first 12 MFCC were calculated using Hamming windows of 25 ms with a 10 ms frame shift. The parameterisation was calculated using the *Hidden Markov Models Toolkit* [27]. After that, similarly to (2) and (3), the mean of each MFCC coefficient was computed along the utterance.

Incorporation of prosodic features

The use of prosodic features in emotion recognition has already been studied and discussed extensively [8, 14, 28]. As in almost all works, here the classic methods to calculate *Energy* and pitch (F_0) along the sentence were used [29]. 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 [16].

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Robust Emotion Recognition using Bio-inspired Features"
Anales de la XV Reunión de Procesamiento de la Información y Control, pp. 94-99, sep, 2013.

Table 1: Classification rate for clean signals. Accuracy is computed with validation data.

Feature Vector	*12MFCC*	*30MLS*	*12PCA-MLS*	*71MLSa*	*12PCA-MLSa*
N_H^*	70	60	45	20	100
Acc [%]	63.17	56.83	56.51	48.89	48.73
Feature Vector	*12MFCC+P*	*30MLS+P*	*12PCA-MLS+P*	*71MLSa+P*	*12PCA-MLSa+P*
N_H^*	90	65	65	55	40
Acc [%]	65.71	59.37	59.21	56.35	**66.19**

Neural classifier

In this work, the standard MLP is used as classifier in all the experiments because this is widely known technique in emotion recognition [7, 14]. MLP is a class of artificial neural network and it consists of a set of process units (simple perceptrons) arranged in layers [30]. In the MLP, the nodes are fully connected between layers without connections 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. 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* [31] was used.

For every feature vector, a preliminar exploration to reach the best configuration 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^*. In all cases, the output layer has seven neurons. Network training was stopped when it reached the generalization point with test data [30], and so, its test rate and trained network were kept. The trained network that achieved the best test results (average over all partitions) was evaluated on clean and noisy validation data.

3 Experiments and results

In this section, the robustness of the proposed feature extraction methods is evaluated. Firstly, the experiments with clean signals are presented. Then, the experiments with noisy signals are discussed. The MLP classifiers trained with clean signals were evaluated considering signals artificially contaminated with additive white noise[1]. Noisy signals were generated using several signal-to-noise ratios (SNR).

Features were arranged in vectors with and without prosodic information. In order to present the experiments and results in a more readable form, the following notation is introduced:

- *12MFCC*: 12 mean MFCC 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.

[1] Using Matlab 7.8.0.347.

The aim of selecting 12 PCA coefficients was to keep the same dimensionality that MFCC coefficients. The features with prosodic values are pointed out using "+ *P*" in notation.

For all the experiments, all the vectors were normalised. Firstly, the maximum and the minimum values (for each dimension) from the training set were extracted. After that, these values were used to normalise the training vectors, clean validation vectors and all sets of noisy validation vectors.

In order to avoid biased estimates of recognition error, a cross-validation method was performed. In the classification experiments, ten data partitions were generated by randomly pick 80% for *training*. The remaining 20% was left for *validation*. The training data in turn was randomly separated in 60% for training and 20% for the generalization test. This is the same setup defined in [16] and thus, the results are directly comparable.

Emotional Speech Corpus

In this work we used a well-known acted database, the *EmoDB*, [32] because it is freely accessible[2] and it was used in several studies [8, 16, 33]. The corpus, consisting of 535 utterances, includes sentences performed under six plain emotions, and sentences in neutral emotional state. The distribution 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). Almost all works did not address this issue, obtaining biased and hardly comparable results. In the same way as in [16], the dataset was balanced by equalizing the size of the classes. The same number of samples for all classes in each partition were randomly selected (46 x 7 = 322 utterances). Each utterance is an unique training or test pattern in a data partition, its transcription is ignored and its label is the name of the emotional class to which it belongs.

3.1 Experiments without noise

The different features without noise were evaluated to set up a reference. We decided to evaluate the different feature vectors with and without prosodic values, then their usefulness in complementing the other features can be analised. Table 1 shows the classification accuracy for all considered features, while the rows N_H^* exhibit the

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

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Robust Emotion Recognition using Bio-inspired Features"
Anales de la XV Reunión de Procesamiento de la Información y Control, pp. 94-99, sep, 2013.

Table 2: Performance of features under noise conditions. Accuracy in [%].

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

Table 3: Robust classification for features with prosody. Accuracy in [%].

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

number of neurons in the hidden layer for the best configuration. As can be seen, results for *12MFCC* are good, showing a significant improvement when prosodic values are used. For vectors based on MLS a satisfactory performance is obtained, and better results are reached when prosodic information is added. In these results can be noticed that the PCA is very useful for dimensional reduction of MLS features. The *30MLS* and the *12PCA-MLS* achieve a similar result (with and without prosody). Bio-inspired features (MLSa), as much *71MLSa* as *12PCA-MLSa*, exhibit a poorer and similar behaviour. However, once more the PCA was good to reduce dimensionality retaining discriminative information. Moreover, when prosodic information was used, the accuracy was improved more than 8 % (*71MLSa+P*). The best accuracy is reached with *12PCA-MLSa+P*, improving almost 18 % (absolute) the classification rate achieved with *12PCA-MLSa*. It points out promising results when the novel MLSa information and prosodic values are used together for the emotion recognition task. Furthermore, similar results were reached using PCA for dimensional reduction, keeping (at the most) the same dimensionality of MFCC.

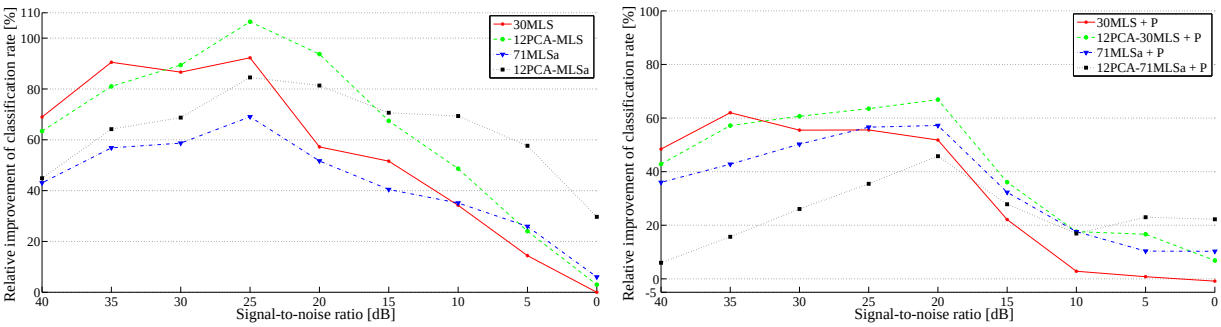
3.2 Experiments with noise

For experiments with noise, we used the MLPs trained with clean signals. We selected the MLPs that have obtained the highest scores (in each case, with and without prosody).

In Table 2, the classification results using noisy signals are presented. In the first row, signal-to-noise ratios are indicated and the remaining rows show the performance of every feature vector. The best performances are emphasized in bold. Similar to previous works, the *12MFCC* works fine with clean signals but its behaviour degrades abruptly in noise presence, approximately 29 % from ∞ dB to 40 dB. As can be observed in the table, the sets of proposed features reached a noticeable improvement in classification rates, respect to MFCC. For example, it can be perceived an increment of more than 23 % (from 34.29 to 57.94) with 40 dB and an increment of more than 27 % (from 30.16 to 57.46) with 35 dB. In $20-30$ dB range, the PCA extracted from MLS is the best to reduce noise effects, whereas it allows a dimensional reduction of 60 % (from 30 to 12 coefficients). Therefore, the 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. For low signal-to-noise ratio ($[5-15]$ dB), the PCA extracted from MLSa obtained a classification enhancement about 10 % and more (from 16.51 to 26.03, from 17.62 to 29.84, and from 20.00 to 34.13). Meanwhile its improvement for 0 dB is 4.73 %. The proposed features based on MLS exhibit the best performances in conditions of high SNR (up to 20 dB) whereas the bio-inspired parameterisation shows a robust behaviour against high presence of noise. It can be observed significant improvements when the dimensionality of features is reduced using PCA. These results would suggest that these representations allow to keep in a separate manner the signal information and the noise. Moreover, the relevant information can be obtained even in the presence of higher levels of noise using the nonlinear filter bank of the auditory model.

In Table 3, the results for the feature vectors with prosodic information are presented. As it can be observed in the second column, prosodic information improves the classification for all representations using clean signals (more than 17% in *12PCA-MLSa* case). However, this is not always true under noisy conditions and it could be attributable to the non-robust method used for prosodic computation. In spite of this, a similar analysis to the previous cases can be done and results show an equivalent behaviour. In this way, the parameterisations based on MLS obtained the best performances for high SNR whereas the parameterisations extracted from the bio-inspired information presents a robust behaviour against high presence of noise. Once again, the PCA represents an important alternative to provide robust behaviour and a lower dimensionality in the input vectors for the clas-

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner; "Robust Emotion Recognition using Bio-inspired Features"
Anales de la XV Reunión de Procesamiento de la Información y Control, pp. 94-99, sep, 2013.



(a) Improvements relative to *12MFCC*, without prosody.

(b) Improvements relative to *12MFCC + P*, with prosody.

Figure 1: Improvement of classification rate.

sifier. Here, it can be noticed an absolute improvement about 20 % in $25 - 40$ dB range and a relative classification improvement about 22 % for 0 dB.

Additional information is showed graphically in Figures 1(a) and 1(b). In both can be observed the relative improvement of classification rate with regard to *12MFCC* and *12MFCC + P*, respectively. As can be noticed in Figure 1(a), there are significant improvements in almost all cases. For example, the proposed features improved the *12MFCC* more than 40 % up to 15 dB, reaching a maximum over 100 % of relative improvement. For low signal-to-noise ratios ($[5 - 15]$ dB), a relative classification enhancement about $60 - 70$ % is obtained. Although performance for all features decay to 0 dB, *12PCA-MLSa* obtains 30 % of relative improvement (from 16.03 to 20.79) with 0 dB SNR. As was mentioned above, Figure 1(b) presents the results for feature vectors with prosody. It is important to note that the incorporation of prosodic information is more beneficial to the reference features (*12MFCC + P*). The proposed features show a superior performance, although the gap of relative improvement decreases with respect to the previous figure. In both figures can be observed that proposed parameterisations have evidenced a good performance under noisy conditions.

As discussed in previous work, the most relevant information for emotion recognition was found between 0 and 1200 Hz. Given the new results, the good behaviour can be attributable 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. On the other hand, it also may be due to the filters and approximations that simulate the remainder behaviour of the early auditory system. Moreover, the PCA keeps relevant characteristics of emotional signals disregarding the noise information while it perform a dimensional reduction of the input vector.

4 Conclusions

In this paper we introduced a new set of bio-inspired features (MLSa) based on a time-frequency analysis computed with an auditory model. The MLSa, MLS and others lower dimensional vectors (computing through the PCA) from the proposed features have been used in emotion classification with noisy signals using MLP.

The MLS and MLSa features, and these combined with prosody, improved the accuracy under additive white noise conditions. Furthermore, PCA coefficients extracted from MLS and MLSa allowed to improve the results while low dimensional data is used. As in previous works, the MFCC showed a good performance with clean signals but the proposed features always improved the accuracy under noise conditions. Features based on spectral information were very good for high SNR, whereas features computed using the auditory model were more robust for low SNR.

In future work the method will be tested under noise condition using matched/mismatched/multi condition training. In addition, the proposed parameterisation could be used to improve the performance of hierarchical classifiers. Furthermore, we will explore the performance of classifiers with another type of non-stationary noises.

5 Acknowledgements

The authors wish to thank to *ANPCyT* and *Universidad Nacional de Litoral* (with PAE 37122, PACT 2011 #58, CAI+D 2011 #58-511) and *CONICET*, for their support.

REFERENCES

[1] C. Clavel, I. Vasilescu, L. Devillers, G. Richard, and T. Ehrette, "Fear-type emotion recognition for future audio-based surveillance systems," *Speech Communication*, vol. 50, no. 6, pp. 487–503, 2008.

[2] D. Tacconi, O. Mayora, P. Lukowicz, B. Arnrich, C. Setz, G. Tröster, and C. Haring, "Activity and emotion recognition to support early diagnosis of psychiatric diseases," in *Proc. of 2nd Int. Conf. on Pervasive Computing Technologies for Healthcare*, Finland, Feb. 2008, pp. 100–102.

[3] S. Yildirim, S. Narayanan, and A. Potamianos, "Detecting emotional state of a child in a conversational computer game," *Computer Speech & Language*, vol. 25, no. 1, pp. 29 – 44, 2011.

[4] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 597–607, 2012.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner, "Robust Emotion Recognition using Bio-inspired Features"
Anales de la XV Reunión de Procesamiento de la Información y Control, pp. 94-99, sep, 2013.

[5] K. Schindler, L. Van Gool, and B. de Gelder, "Recognizing emotions expressed by body pose: A biologically inspired neural model," *Neural Networks*, vol. 21, no. 9, pp. 1238–1246, 2008.

[6] A. Batliner, S. Steidl, B. Schuller, D. Seppi, T. Vogt, J. Wagner, L. Devillers, L. Vidrascu, V. Aharonson, L. Kessous, and N. Amir, "Whodunnit - Searching for the most important feature types signalling emotion-related user states in speech," *Computer Speech & Language*, vol. 25, no. 1, pp. 4–28, 2011.

[7] M. El Ayadi, M. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572–587, 2011.

[8] M. Borchert and A. Dusterhoft, "Emotions in speech - experiments with prosody and quality features in speech for use in categorical and dimensional emotion recognition environments," in *Proc. of IEEE Int. Conf. on Natural Language Proc. and Knowledge Engineering*, Oct. 2005, pp. 147–151.

[9] Y.-L. Lin and G. Wei, "Speech emotion recognition based on HMM and SVM," in *Proc. of Int. Conf. on Machine Learning and Cybernetics*, vol. 8, Aug. 2005, pp. 4898–4901.

[10] J. Wagner, T. Vogt, and E. André, "A Systematic Comparison of Different HMM Designs for Emotion Recognition from Acted and Spontaneous Speech," in *Affective Computing and Intelligent Interaction*, ser. LNCS. Springer Berlin Heidelberg, 2007, vol. 4738, pp. 114–125.

[11] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Tran. on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39–58, Jan. 2009.

[12] D. Morrison, R. Wang, and L. C. D. Silva, "Ensemble methods for spoken emotion recognition in call-centres," *Speech Communication*, vol. 49, no. 2, pp. 98–112, 2007.

[13] L. Fu, X. Mao, and L. Chen, "Speaker independent emotion recognition based on SVM/HMMs fusion system," in *Proc. of Int. Conf. on Audio, Language and Image Processing*, Jul. 2008, pp. 61–65.

[14] B. Schuller, G. Rigoll, and M. Lang, "Speech emotion recognition combining acoustic features and linguistic information in a hybrid support vector machine-belief network architecture," in *Proc. of IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, May 2004, pp. I–577–80.

[15] Z. Xiao, E. Dellandréa, W. Dou, and L. Chen, "Recognition of emotions in speech by a hierarchical approach," in *Proc. of Int. Conf. on Affective Computing and Intelligent Interaction*, Sep. 2009, pp. 312–319.

[16] E. M. Albornoz, D. H. Milone, and H. L. Rufiner, "Spoken emotion recognition using hierarchical classifiers," *Computer Speech & Language*, vol. 25, no. 3, pp. 556–570, 2011.

[17] B. Schuller, S. Steidl, A. Batliner, F. Burkhardt, L. Devillers, C. Müller, and S. Narayanan, "Paralinguistics in speech and language - state-of-the-art and the challenge," *Computer Speech & Language*, vol. 27, no. 1, pp. 4–39, 2013, Special issue on Paralinguistics in Naturalistic Speech and Language.

[18] D. Sztahó, V. Imre, and K. Vicsi, "Automatic classification of emotions in spontaneous speech," in *Analysis of Verbal and Nonverbal Communication and Enactment. The Processing Issues*, ser. LNCS. Springer Berlin Heidelberg, 2011, vol. 6800, pp. 229–239.

[19] A. Tawari and M. Trivedi, "Speech emotion analysis in noisy real-world environment," in *Proc. of 20th Int. Conf. on Pattern Recognition*, Aug. 2010, pp. 4605–4608.

[20] Z. Han, S. Lun, and J. Wang, "A study on speech emotion recognition based on CCBC and neural network," in *Proc. of Int. Conf. on Computer Science and Electronics Engineering*, vol. 2, Mar. 2012, pp. 144–147.

[21] B. Schuller, D. Seppi, A. Batliner, A. Maier, and S. Steidl, "Towards more reality in the recognition of emotional speech," in *Proc. of IEEE Int. Conf. on Acoustics, Speech and Signal Proc.*, vol. 4, Apr. 2007, pp. IV–941–IV–944.

[22] S. A. Shamma, R. S. Chadwick, W. J. Wilbur, K. A. Morrish, and J. Rinzel, "A biophysical model of cochlear processing: Intensity dependence of pure tone responses," *The Journal of the Acoustical Society of America*, vol. 80, no. 1, pp. 133–145, 1986.

[23] X. Yang, K. Wang, and S. A. Shamma, "Auditory representations of acoustic signals," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 824–839, Mar. 1992.

[24] T. Chi, P. Ru, and S. A. Shamma, "Multiresolution spectrotemporal analysis of complex sounds," *Journal of the Acoustical Society of America*, vol. 118, no. 2, pp. 887–906, 2005.

[25] N. Mesgarani and S. Shamma, "Denoising in the domain of spectrotemporal modulations," *EURASIP Journal on Audio, Speech, and Music Processing*, vol. 2007, no. 3, pp. 1–8, Jul. 2007.

[26] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer, Aug. 2006.

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

[28] J. Adell Mercado, A. Bonafonte Cávez, and D. Escudero Mancebo, "Analysis of prosodic features: towards modelling of emotional and pragmatic attributes of speech," *Procesamiento de Lenguaje Natural*, no. 35, pp. 277–283, 2005.

[29] J. R. Deller Jr., J. G. Proakis, and J. H. Hansen, *Discrete-Time Processing of Speech Signals*. Upper Saddle River, NJ, USA: Prentice Hall PTR, 1993.

[30] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Prentice Hall, Jul. 1998.

[31] A. Zell, G. Mamier, M. Vogt, N. Mache, R. Hubner, S. Doring, K.-U. Herrmann, T. Soyez, M. Schmalzl, T. Sommer, A. Hatzigeorgiou, D. Posselt, T. Schreiner, B. Kett, and G. Clemente, *SNNS (Stuttgart Neural Network Simulator)*, Germany, 1998, User Manual Version.

[32] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A Database of German Emotional Speech," in *Proc. of 9th European Conf. on Speech Communication and Technology*, Sep. 2005, pp. 1517–1520.

[33] B. Yang and M. Lugger, "Emotion recognition from speech signals using new harmony features," *Signal Processing*, vol. 90, no. 5, pp. 1415–1423, 2010, Special Section on Statistical Signal & Array Processing.

sinc(*i*) Research Center for Signals, Systems and Computational Intelligence (fich.unl.edu.ar/sinc)
E. M. Albornoz, D. H. Milone & H. L. Rufiner, "Robust Emotion Recognition using Bio-inspired Features"
Anales de la XV Reunión de Procesamiento de la Información y Control, pp. 94-99, sep, 2013.