Non-linear Dynamics Characterization from Wavelet Packet Transform for Automatic Recognition of Emotional Speech

J.C. Vásquez-Correa[1*], J.R. Orozco-Arroyave[1,2], J.D. Arias-Londoño[3], J.F. Vargas-Bonilla[1], and Elmar Nöth[2]

[1] Electronics and Telecommunications engineering Department. Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia.
[2] Pattern Recognition Lab., Friedrich-Alexander-Unversität, Erlangen-Nürnberg, Germany.
[3] Computer engineering Department. Universidad de Antioquia UdeA, Calle 70 No. 52-21, Medellín, Colombia.
[*] Corresponding author: jcamilo.vasquez@udea.edu.co

Abstract. We propose a new set of features based on non-linear dynamics measures obtained from the wavelet packet transform for the automatic recognition of "fear-type" emotions in speech. The experiments are carried out using three different databases with a Gaussian Mixture Model for classification. The results indicate that the proposed approach is promising for modeling "fear-type" emotions in speech.

Keywords: Non-linear Dynamics, Non-linear speech processing, Speech emotion recognition, Wavelet Packet Transform

1 Introduction

Speech is the main process of communication between humans. This fact has motivated researches to use it as a mechanism of interaction between humans and computers. The challenge now is not only to recognize the words and sentences but also the paralinguistic aspects of speech such as emotions and personality of the speaker. In the last few years the interest of the research community has been focused on the detection of "fear-type" emotions such as anger, disgust, fear, and desperation, which appear in abnormal situations when the human integrity is at risk [1]. One of the main aims of speech analysis is to find suitable speech features to represent the emotional state of a speaker. In related works, the characterization has been focused on prosodic features, spectral and cepstral features such as Mel Frequency Cepstral Coefficients (MFCC), and voice quality features such as noise measures [1]. In [2] the authors use Berlin [3], and enterface05 [4] databases for emotion recognition. They use MFCC joint to their first and second derivatives, and perform the classification using a Deep Neural Network with a Hidden Markov Model (DNN-HMM). The reported accuracies are 77.92%, and 53.89% for Berlin, and enterface05 databases, respectively. In [5]

the authors use enterface05 [4], and FAU Aibo [6] databases for emotion recognition. They use acoustic features related to MFCC, energy, and fundamental frequency, and propose a method based on least square regression (LSR) for recognition. The reported accuracies are 69.33% in enterface05, and 60.50% in Aibo database, respectively.

On the other hand, the use of Non-Linear Dynamics (NLD) measures in speech processing tasks has been increased in the last years. In [7] the authors perform experiments in three different databases, including the Emotional prosody speech and transcripts of the Linguistic Data Consortium (LDC) [8], the Berlin database [3], and the Polish emotional speech database [9]. The authors characterize the speech recordings using features related to NLD and perform the classification of different emotional states using an artificial neural network. The reported accuracies are 80.75%, 75.40%, and 72.78% for each database, respectively. In [10] the authors use Berlin, and SUSAS [11] databases to evaluate the representation capability of the Hurst exponent (HE) obtained from the Discrete Wavelet Transform (DWT) to recognize different emotions and to detect stress from speech. The author perform segmentation between voiced and unvoiced segments, and calculate the features only for voiced segments. The speech signals are modeled using a Gaussian Mixture Model (GMM), and the reported accuracies are 68.1%, and 64%, for the Berlin and SUSAS databases, respectively. In [12] the authors use SUSAS [11] database for automatic stress recognition in speech. They use features related to energy and entropy obtained from Wavelet Packet Transform (WPT). Automatic recognition is performed by means of a Linear Discriminant Analysis (LDA), and the reported accuracy is about 91%.

In previous works, we calculate acoustic features obtained from WPT [13]. In this paper, we propose a new set of features related to NLD obtained from WPT for fear-type emotion recognition in speech signals. WPT provides a time-frequency representation in different resolutions and the NLD features are calculated on each decomposed band. The features are calculated on speech recordings of three different databases very used in emotion recognition: (i) Berlin [3], (ii) GVEESS [14], and (iii) enterface05 [4]. Classification is performed using a GMM derived from a Universal Background Model (GMM-UBM). The rest of paper is distributed as follows: section 2 contains the description about the characterization and classification processes. Section 3 describes the experimental framework, the databases, and the obtained results. Finally, section 4 includes the conclusions derived from this study.

2 Materials and methods

Figure 1 shows the general scheme of the proposed methodology. It consists of four stages. First the voiced and unvoiced segments of speech are separated using the software Praat [15] in order to analyze features estimated from each one of them. Second the wavelet decomposition is performed on each segment separately. Third each decomposed band is characterized separately as follows: for the

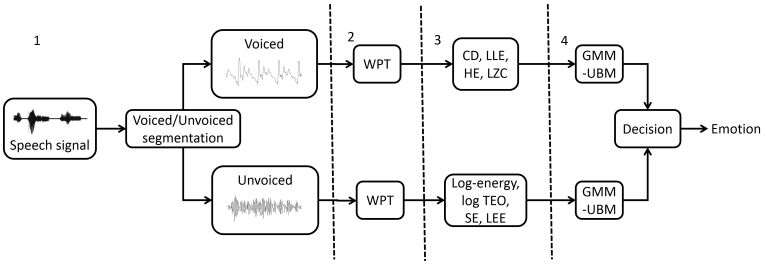

Fig. 1. General scheme of the methodology

voiced segments four NLD features are calculated including Correlation Dimension (CD), Largest Lyapunov exponent (LLE), HE, and Lempel-Ziv complexity (LZC); and for the case of unvoiced segments another four features are estimated including the log energy, the log energy derived from Teager Energy Operator (TEO), the Shannon Entropy (SE), and the Log Energy Entropy (LEE). The difference in the features estimated both for voiced and unvoiced segments is based on the fact that features estimated for voiced segments are related to perturbation of the fundamental frequency, and the excitation source [10]. These features cannot be estimated for unvoiced segments. Finally, the fourth stage of the methodology includes the GMM-UBM modeling. The decision of which emotion is present on each recording is taken by the combination of the posterior probabilities produced by the classifiers applied on the voiced and unvoiced feature vectors.

2.1 Feature estimation

Taken's embedding and phase space: the NLD analysis begins with the reconstruction of the phase space of the speech signal according to the embedding process [7]. A time series $x(i)$ $i = 1, 2, \cdots N_m$, can be represented in a new space which is defined as $X[k] = \{\mathbf{x}[k], \mathbf{x}[k+\tau], \mathbf{x}[k+2\tau], \mathbf{x}[k+(m-1)\tau]\}$. $N_m = N - (m-1)\tau$ is the reconstructed vector length, τ is the time delay, and m is the embedding dimension.

Correlation Dimension (CD): this feature allows the estimation of the exact space that is occupied by the reconstructed vector in the phase space. It is an indicator about the complexity and dimensionality of speech signal [7].

Largest Lyapunov Exponent (LLE): this measure quantifies the exponential divergence of neighbor trajectories in a phase space. This measure provides an indicator of the aperiodicity of a speech signal [7].

Hurst Exponent (HE): this feature expresses the long term dependence of a time series. HE is defined according to the asymptotic behaviour of the rescaled range of a time series as a function of a time interval. HE can be used to represent the emotional state of speech according to the arousal level of the signal [10]:

Lempel Ziv Complexity (LZC): this feature establishes a measure about the degree of disorder of spatio-temporal patterns in a time series. The LZC reflects the rate of new patterns in the sequence; and ranges from 0 (deterministic sequence) to 1 (random sequence). LZC distribution shows values nearer to 1 for fear and anger speech, than in case of neutral speech [7].

log-Energy: this is a classical feature to characterize emotional speech. It is calculated according to the equation 1 [16].

$$E(k) = log\left[\frac{\sum_{l=1}^{N_k} x(l)^2}{N_k}\right] \tag{1}$$

log-Energy of TEO: the TEO was developed in order to measure the changes in speech signal produced under stress. The TEO of a signal is calculated as $TEO(x) = x(k)x(k)^* - x(k+1)x(k-1)$. Finally, the log Energy of TEO is calculated according to the equation 2 [16].

$$E_{TEO}(k) = log\left[\frac{\sum_{l=1}^{N_k} |x(l)x(l)^* - x(l+1)x(l-1)|}{N_k}\right] \tag{2}$$

Shannon and log energy entropy: entropy describes the complexity of a system. In this paper we estimate the Shannon, and log energy entropy. These features are calculated using the equations 3 and 4, respectively [12]. n is the number of bins used to estimate the probability density function of the wavelet decomposition.

$$H1(k) = -\sum_{l=1}^{n} x(l)^2 log(x(l)^2) \tag{3}$$

$$H2(k) = -\sum_{l=1}^{n} log(x(l)^2) \tag{4}$$

2.2 classification

The features extracted from voiced and unvoiced segments were classified separately using a GMM adapted from a UBM, using Maximum A Posterior (MAP) to derive a specific GMM from the UBM [17]. A GMM can be defined as a probabilistic model represented by the sum of several multivariate Gaussian components. The model is expressed according to its probability density function.

$$p(\mathbf{x}|\Theta) = \sum_{k=1}^{M} P_k \mathcal{N}(\mathbf{x}|\mu_k, \mathbf{\Sigma}_k) \tag{5}$$

Where M is the number of Gaussian components, P_k is the prior probability (mixing weight), and $\mathcal{N}$ is a multivariate Gaussian density function. The UBM

is trained using the Expectation Maximization (EM) algorithm [17] using a population of all classes. Then the specific GMM for each class is adapted using the MAP method. Finally, given a sample $X = \mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_T$, where $\mathbf{x}_i$ is the feature vector extracted from the frame i, the decision about to which class belongs each speech sample is taken evaluating the maximum Log-Likelihood according to the equation 6.

$$LL(X, \Theta) = \sum_{t=1}^{T} log(p(\mathbf{x}_t|\Theta)) \tag{6}$$

In this work, the posterior probability obtained from the GMM based on voiced segments is combined with the obtained from the GMM model based on unvoiced segments. The new probability LL_{comb} is expressed as $LL_{comb} = \alpha LL(X, \Theta_{Voiced}) + (1-\alpha)LL(X, \Theta_{Unvoiced})$. Where α denotes weight coefficient.

3 Experimental framework and results

3.1 experimental setup

All experiments were performed using leave one group speaker out cross-validation (LOGSO-CV), with different numbers of Gaussian components for classifier (from 2 to 8) with diagonal covariance matrix. The value of α was optimized through a grid search with $0 < \alpha < 1$. The selection criteria was based on the obtained accuracy on the test set. The length of frames is selected according to the sample frequency in order to guarantee enough number of points for the wavelet decomposition. Frames of 1764 samples are selected, which is equivalent to $40ms$ in cases when $fs = 44100Hz$, and $110ms$ when $fs = 16000Hz$ [13]. In both cases an overlapping percentage of 50% is used. All of the coefficients from the second and third level of WPT are considered. Daubechies3 is used as the mother wavelet. The experiments were carried out to recognize different "fear-type" emotions in speech, such as anger, disgust, fear and desperation.

3.2 Databases

- Berlin emotional database [3]: it contains 534 voice recordings produced by 10 speakers who acted 7 different emotions. The recordings were sampled at 16KHz. In this paper three of the seven emotions of the database are considered for the automatic recognition: anger, disgust, and fear.
- Geneva Vocal Emotion Stimulus Set (GVEESS) [14]: it contains 224 recordings of 12 speakers who acted 14 emotions. The recordings were sampled at 44.1KHz. In this paper four of the 14 emotions of the database are considered for the automatic recognition: anger, disgust, fear, and desperation.
- enterface05 database [4]: this database contains 1317 audio-visual recordings with 6 emotions produced by 42 speakers. In this paper three of the six emotions of the database are considered for the automatic recognition: anger, disgust, and fear. In this database, each subject was instructed to listen to

six successive short stories. After each story the subject had to react to the situation by speaking predefined phrases that fit the short story.

3.3 Results

Tables 1 and 2 contain the accuracies obtained with the features extracted from voiced and unvoiced segments, respectively. The number of Gaussian components is also indicated in the tables. In Table 1 the highest accuracies are obtained with the LZC. Table 2 shows that the best results are reached using the combination of all features, specially in GVEESS, and enterface05 databases. Note also that in this case it is possible to achieve similar results than in the case of features estimated only for voiced segments, or features estimated without the segmentation process as in related works [2, 7].

Table 1. Emotion recognition accuracies for features estimated from voiced segments

	GVEESS		Berlin		enterface05	
Features	Accuracy	M	Accuracy	M	Accuracy	M
DC	57.1±14.6	4	62.7±13.9	5	47.6±3.8	4
LLE	68.0±16.2	4	67.6±8.1	5	52.1±4.9	6
HE	68.1±28.0	4	67.6±8.1	5	52.0±4.9	6
LZC	**82.0±11.3**	**3**	**78.3±9.9**	**3**	**54.0±7.3**	**4**
All	65.0±21.2	5	79.0±10.0	4	51.1±8.0	5

Table 2. Emotion recognition accuracies for features estimated from unvoiced segments

	GVEESS		Berlin		enterface05	
Features	Accuracy	M	Accuracy	M	Accuracy	M
Log Energy	93.4±9.8	4	64.7±11.1	4	46.9±4.4	5
Log Energy TEO	93.1±8.8	6	60.8±10.3	5	54.2±4.9	5
SE	93.4±9.8	4	71.0±12.7	4	53.7±5.8	4
LEE	92.3±10.3	5	**77.2±10.9**	**4**	57.0±4.1	6
All	**99.0±2.5**	**6**	69.13±16.0	6	**63.1±15.7**	**3**

The posterior probabilities obtained with the voiced and unvoiced features that reached highest accuracies are combined. The results are shown in Figure 2. Note that the combination of posterior probabilities provides better accuracy rate than the separately classification for voiced and unvoiced segments. In Figure 2 each recognized emotion is indicated in the x axis. The bars indicate the accuracies obtained on each emotion, i.e. in the Berlin database the bars on Disgust indicate that 24% of the recordings labeled as Disgust where wrongly recognized by the system as Fear (bar in black). Also 13% of the Disgust recordings where wrongly recognized as Anger. Finally, 63% of the Disgust recordings were correctly recognized by the system as Disgust. Note that the highest accuracy in all of the three databases is obtained with the recordings labeled as Anger. This result shows that the proposed feature set is able to model the fast

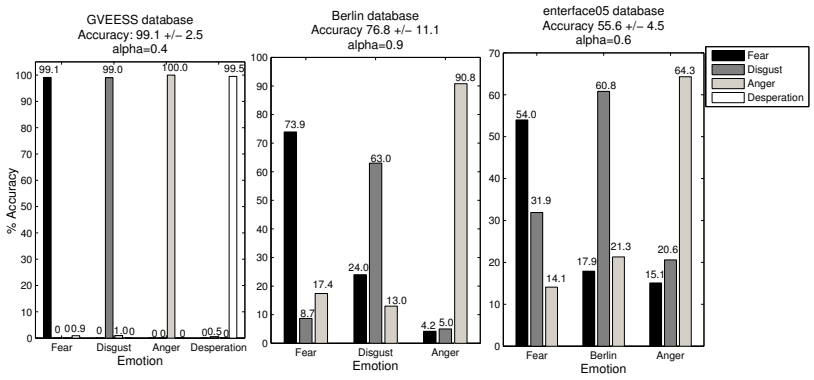

Fig. 2. Results of combination of posterior probabilities for voiced and unvoiced segments

air flow in vocal tract produced by anger in speech, which causes vortices located near the false vocal folds providing additional excitation signals other than the pitch [18].

4 Conclusion

A total of four NLD features obtained from the WPT are extracted from speech signals to perform the automatic recognition of fear-type emotions. The voiced and unvoiced segments of each recording are characterized separately.

The results show that LZC evaluated from wavelet decomposition in voice segments provides a good representation of emotional content in speech signal relative to the other NLD measures estimated from WPT. We found also that features derived from energy and entropy content of unvoiced segments are suitable for the characterization of emotional speech. The obtained results are similar to those obtained in related works when classical features are used, indicating that features related to NLD are useful to represent the emotional content in speech, and must be used to characterization of emotional speech.

The evaluation of the proposed features in speech recordings with non-controlled scenarios such as phone channels, signals with background noise, and non-acted emotions needs to be addressed in future work.

Acknowledgment

Juan Camilo Vásquez Correa is granted by the program of young researchers and innovators 2014, financed by COLCIENCIAS. Juan Rafael Orozco Arroyave is under grants of Convocatoria 528 para estudios de doctorado en Colombia 2011 financed by COLCIENCIAS. The authors thank to CODI, estrategia de sostenibilidad 2014-2015 from Universidad de Antioquia for the support for the development of this work.

References

1. Schuller, B., Batliner, A., Steidl, S., Seppi, D.: Recognising realistic emotions and affect in speech: State of the art and lessons learnt from the first challenge. Speech Communication **53**(9-10) (November 2011) 1062–1087
2. Li, L., Zhao, Y., Jiang, D., Zhang, Y., Wang, F., Gonzalez, I., Valentin, E., Sahli, H.: Hybrid deep neural network-hidden markov model (dnn-hmm) based speech emotion recognition. In: Proceedings of the 2013 Humaine Association Conference on Affective Computing and Intelligent Interaction. (2013) 312–317
3. Burkhardt, F., Paeschke, A., Rolfes, M., Sendlmeier, W., Weiss, B.: A database of german emotional speech. Proc of the INTERSPEECH 2005 (2005) 1517–1520
4. Martin, O., Kotsia, I., Macq, B., Pitas, I.: The enterface'05 audio-visual emotion database. In: Proceedings of the 22Nd International Conference on Data Engineering Workshops. ICDEW '06 (2006) 8–15
5. Zheng, W., Xin, M., Wang, X., Wang, B.: A novel speech emotion recognition method via incomplete sparse least square regression. Signal Processing Letters, IEEE **21**(5) (May 2014) 569–572
6. Steidl, S.: Automatic Classification of Emotion-Related User States in Spontaneous Children s Speech. (2009)
7. Henríquez, P., Alonso, J.B., Ferrer, M.A., Travieso, C.M., Orozco-Arroyave, J.R.: Nonlinear dynamics characterization of emotional speech. Neurocomputing **132** (2014) 126–135
8. Liberman, M., Davis, K., Grossman, M., Martey, N., Bell, J.: Emotional prosody speech and transcripts ldc2002s28 (2002)
9. Staroniewicz, P., Majewski, W.: Polish emotional speech database - recording and preliminary validation. In: COST 2102 Conference (Prague). Volume 5641 of Lecture Notes in Computer Science., Springer (2008) 42–49
10. Zao, L., Cavalcante, D., Coelho, R.: Time-frequency feature and ams-gmm mask for acoustic emotion classification. Signal Processing Letters, IEEE **21**(5) (May 2014) 620–624
11. Hansen, J.H.L., Bou-Ghazale, S.E.: Getting started with susas: a speech under simulated and actual stress database. In: EUROSPEECH, ISCA (1997)
12. Bt Johari, N., Hariharan, M., Saidatul, A., Yaacob, S.: Multistyle classification of speech under stress using wavelet packet energy and entropy features. In: Sustainable Utilization and Development in Engineering and Technology (STUDENT), 2011 IEEE Conference on. (Oct 2011) 74–78
13. Vasquez-Correa, J., Garcia, N., Vargas-Bonilla, J., Orozco-Arroyave, J., Arias-Londoño, J., Quintero, O.: Evaluation of wavelet measures on automatic detection of emotion in noisy and telephony speech signals. In: Security Technology (ICCST), 2014 International Carnahan Conference on. (Oct 2014) 1–6
14. Banse, R., Scherer, K.: Acoustic profiles in vocal emotion expression. Journal of Personality and Social Psychology **70** (1996) 614–636
15. Boersma, P., Weenik, D.: Praat: a system for doing phonetics by computer. report of the institute of phonetic sciences of the university of amsterdam. (1996)
16. Kandali, A.B., Routray, A., Basu, T.K.: Vocal emotion recognition in five native languages of Assam using new wavelet features. International Journal of Speech Technology **12**(1) (October 2009) 1–13
17. Reynolds, D.A., Quatieri, T.F., Dunn, R.B.: Speaker verification using adapted gaussian mixture models. Digital Signal Processing **10**(1-3) (2000) 19–41
18. Ververidis, D., Kotropoulos, C.: Emotional speech recognition: Resources, features, and methods. Speech Communication **48**(9) (2006) 1162–1181