Sensors & Transducers, Vol. 204, Issue 9, September 2016, pp. 45-57

Sensors & Transducers

http://www.sensorsportal.com

Exploring Recurrence Properties of Vowels for Analysis of Emotions in Speech

*** Angela LOMBARDI, Pietro GUCCIONE and Cataldo GUARAGNELLA**

Dipartimento di Ingegneria Elettrica e dell'Informazione - Politecnico di Bari,
via Orabona, 4, 70125 Bari, Italy
* E-mail: angela.lombardi@poliba.it

Received: 4 July 2016 /Accepted: 31 August 2016 /Published: 30 September 2016

Abstract: Speech Emotion Recognition (SER) is a recent field of research that aims at identifying the emotional state of a speaker through a collection of machine learning and pattern recognition techniques. Features based on linear source-filter models have so far characterized emotional content in speech. However, the presence of nonlinear and chaotic phenomena in speech generation have been widely proven in literature. In this work, recurrence properties of vowels are used to describe nonlinear dynamics of speech with different emotional contents. An automatic vowel extraction module has been developed to extract vowel segments from a set of spoken sentences of the publicly available German Berlin Emotional Speech Database (EmoDB). Recurrence Plots (RPs) and Recurrence Quantitative Analysis (RQA) have been used to explore the dynamic behavior of six basic emotions (anger, boredom, fear, happiness, neutral, sadness). Statistical tests have been performed to compare the six groups and check possible differences between them. The results are promising since some RQA measures are able to capture the key aspects of each emotion. *Copyright © 2016 IFSA Publishing, S. L.*

Keywords: Speech Emotion Recognition, Complex systems, Recurrence plot, Recurrence Quantitative Analysis.

1. Introduction

The last decade has seen the rapid and growing development of new algorithms and methods to make the process of human-machine interaction more natural, creating the so-called "Affective Computing" [1]. In many application areas of Artificial Intelligence (e.g. Ambient Assisted Living, Virtual Reality, Smart Recommended System) is required the presence of intelligent agents able to recognize human emotions and process different types of information in order to synthesize empathic reactions.

Among the various ways to detect the emotional state of a user, the employment of some parameters of the speech signal, seems to be one of the most rapid and efficient. Indeed, the presence of a particular affective state is due to the triggering of a series of reactions that take place in the nervous system, which dynamically modify some characteristics of the organs involved in the production of the speech [2].

Speech Emotion Recognition (SER) is a recent field of research that aims at identifying the emotional state of a speaker through a collection of machine learning and pattern recognition techniques. Until now, SER has been used in different application contexts improving the overall performance of the automated systems in which it has been built in: e.g. by detecting the degree of satisfaction of users in the interaction with remote customer-care services, by allowing better communication with students in computer-enhanced learning, by monitoring the stress and attention levels of a driver for in-car board systems and so on (for a complete review see [3-4]).

As a classification problem, a SER system needs a set of features able to optimally reflect the emotional content in speech. According to the existing literature, it is possible to distinguish three main categories of features: prosodic, spectral, and quality-based [5]. Prosodic features such as the fundamental frequency (pitch), the energy of the signal and the rhythm/articulation rate, have been combined with spectral measures (Mel Frequency Cepstral Coefficients (MFCC), Linear Predictor Cepstral Coefficients (LPCC) and formants) in different ways to improve the performances of the classifier [6]. The third category includes acoustic cues related to the shape of glottal pulse signal, its amplitude variation (shimmer) and frequency variation (jitter) [7].

Despite the great variety of classification methods developed for SER applications, still there is no agreement on an optimal set of speech features that can describe and uniquely identify a group of emotional states [3]. This fragmentation of thought is due to several factors. The various sets of features reflect different mechanisms involved in the production of speech sounds but robust theoretical basis about the link between the characteristics of the speech and the emotional state of a speaker does not exist yet [6]. In addition, all the mentioned categories of features are based on a source-filter model [8-9], which represents a simplification of the process of voice production that ignores more complex physiological mechanisms.

Numerous studies carried on since the 1990 s [10-12], have confirmed the presence of non-linear phenomena in speech generation. From these discoveries, new nonlinear tools for speech signal processing have been employed to overcome the limitations imposed by the linear model. In particular, the evidence of the chaotic behavior of some processes involved in the speech production (e.g. turbulent airflow) [13], made the *Chaos Theory* a favored approach for the study of nonlinear dynamics in the system voice.

To describe these dynamics it is necessary to reconstruct the phase space, which is the set of the possible states that the system can take. This approach assumes that the speech signal represents a projection of a higher-dimensional nonlinear dynamical system evolving in time, with unknown characteristics. Embedding techniques can be employed to reconstruct the attractor of the system in the phase space and provide a representation of its trajectories. Afterward, it is possible to describe the dynamic behavior of the system by studying the properties of the embedded attractor: chaotic measures such as Lyapunov exponents, correlation dimension and entropy, have been successfully applied to the analysis of vocal pathologies and speech nonlinearities [14-15].

The behavior of the trajectories of a system in the phase space can also be modeled through the recurrence, a property that quantifies the tendency of a system to return to a state close to the initial one [16]. By exploring similarities between different states at different time epochs, useful information can be provided on the long-term behavior of a system and important aspects about its nature can be revealed. The behavior of the trajectories of a system in the phase space can be easily viewed by means of a recurrence plot (RP). This tool was introduced by Eckmann [17] to facilitate the analysis of the properties for systems with high-dimensional phase space. In contrast to the most of chaotic measures, it is an effective tool even for short and non-stationary data. Recurrence Quantitative Analysis (RQA) [18-19] supplies a quantitative description of the structures contained in a RP through some nonlinear measures. RQA methods have been widely applied in various research fields including biology, astrophysics, engineering, neuroscience, analysis of audio signals and, recently, also for detection and classification of voice disorders [20-22].

In this work we have extended a framework presented in a previous article [23] to explore the recurrence properties of vowel segments taken from a set of spoken sentences of a publicly available database, for six categories of basic emotions (anger, boredom, fear, happiness, neutral, sadness). An automatic vowel extraction module has been built up to extract vowel segments from each sentence; then, their time evolutions have been analyzed by means of the RQA measures. To test the ability of these measures to characterize the different emotional contents, they have been grouped according to the emotion they belong to and statistical tests have been performed to compare the six groups.

The rest of the paper is divided into four sections: theoretical background, general framework, results, discussion and conclusions. In Section 2 theoretical notions on dynamic systems, reconstruction of phase space and recurrence properties are provided; the framework adopted is exposed in details in Section 3; qualitative and quantitative results are shown in Section 4; finally, discussion and conclusions are presented.

2. Theoretical Background

This section provides a general overview of the basic concepts related to the state space reconstruction of a dynamical system and of the main tools used for the analysis of its recurrence properties.

2.1. The Embedding Theorem

The state of a dynamical system is determined by the values of the variables that describe it at a given time. When such system evolves in time, it defines a trajectory in a multidimensional state space, given by the sequence of points that represent all the states of the system. Starting from different initial conditions, a real physical dissipative system tends to evolve in similar ways, so its trajectories converge in a region of the phase space called attractor, which represents the steady state behavior of the system [24].

However, in a real scenario, not all the variables of the system can be inferred and often only a time series $\{u_i\}_{i=1}^{N}$ is available as an output of the system.

Takens demonstrated that it is possible to use time delayed versions of the signal at the output of the system to reconstruct a phase space topologically equivalent to the original one. According to Takens' embedding theorem [25], a state in the reconstructed phase space is given by a m-dimensional time delay embedded vector:

$$\vec{x}_i = \left(u_i, u_{i+\tau}, \ldots, u_{i+(m-1)\tau}\right), \quad (1)$$

where m is the embedding dimension and τ is the time delay.

If $m \geq 2D + 1$, where D is the correlation dimension of the attractor, the original and the reconstructed attractor are diffeomorphically equivalent so the properties of the dynamical system are preserved.

For the embedded parameters estimation, several techniques have been proposed. As an example, the First Local Minimum of Average Mutual Information algorithm [26] can be used to determine when the samples of the time series are independent enough to be useful as coordinates of the time delayed vectors. On the other hand, the false nearest-neighbors algorithm [27] is the method usually employed to estimate the minimum embedding dimension.

2.1. Recurrence Plots

A recurrence plot is a graphical tool that provides a representation of recurrent states of a dynamical system through a two-dimensional square matrix:

$$R_{i,j}(\varepsilon) = \Theta\left(\varepsilon - \left\|\vec{x}_i - \vec{x}_j\right\|\right), \quad i, j = 1, \ldots, N \quad (2)$$

With $\vec{x}_i, \vec{x}_j$ the system state at times i and j, Θ the Heaviside function, ε a threshold for closeness, N is the number of considered states and $\|\bullet\|$ a norm function.

The recurrence matrix contains the value one for all pairs of neighboring states below the threshold ε and zero elsewhere; therefore it allows a quick and effective visual inspection of the dynamic behavior of the system.

The value of the parameter ε must be estimated carefully, as it influences the creation of structures in the plot. In literature, there are some heuristic indications that guide the selection of an appropriate value for such threshold. In general, by choosing ε equal to a few percent of the maximum phase space diameter, a sufficient number of structures in the recurrence plot are preserved, reducing at the same time the presence of artifacts [28].

The resulting plot is symmetric and always exhibits the main diagonal, called line of identity (LOI). Apart for the general RP structure, it is often possible to distinguish small scale structures, which show local (temporal) relationships of the segments of the system trajectory (for a visual reference, see Fig. 4). In details:

- Single isolated points are related to rare states;
- Diagonal lines parallel to the LOI indicate that the evolution of states is similar at different times;
- Vertical lines mark time intervals in which states do not change.

2.1. Recurrence Quantitative Analysis

Several measures of complexity (RQA) have been proposed to obtain an objective quantification of the patterns in a recurrence plot [18-19].

RQA can be divided into three major classes:

1) Measures based on recurrence density. Among these, the simplest measure is the *recurrence rate* (*RR*) defined as:

$$RR(\varepsilon) = \frac{1}{N^2}\sum_{i,j=1}^{N} R_{i,j}(\varepsilon) \quad (3)$$

It is a measure of the density of the recurrence points in the RP.

2) Measures based on the distribution $P(l)$ of lengths l of the diagonal lines. Among these:

- The *determinism* (*DET*) is the ratio of the recurrence points that form diagonal structures (with minimum length l_{min}) to all recurrence points and it is an index of the predictability of a system:

$$DET = \frac{\sum_{l=l_{\min}}^{N} lP(l)}{\sum_{l=1}^{N} lP(l)} \quad (4)$$

- The average diagonal line length (L) is the average time in which two segments of the trajectory move close together:

$$L = \frac{\sum_{l=l_{\min}}^{N} lP(l)}{\sum_{l=1}^{N} P(l)} \quad (5)$$

- The length of the longest diagonal line (L_{max}) found in the RP is related to the exponential divergence of the phase space trajectory:

$$L_{\max} = \left(\{l_i\}_{i=1}^{N_l}\right), \quad (6)$$

- where N_l is the total number of diagonal lines.
- The entropy (ENTR) shows the complexity of the diagonal lines in a RP. It is the Shannon entropy of the probability $p(l)$ to find a diagonal line of length l in the RP:

$$ENTR = -\sum_{l=l_{\min}}^{N} p(l) \ln p(l) \quad (7)$$

- The RATIO, defined as the ratio between DET and RR, combines the advantages of the two categories of measures: it has been proven that it is able to detect some types of transitions in particular dynamics.

3) Measures based on the distribution $P(v)$ of vertical line lengths v. This distribution is used to quantify laminar phases during which the states of a system change very slowly or do not change at all.

- The ratio of recurrence points forming vertical structures longer than v_{min} to all recurrence points of the RP is called laminarity (LAM):

$$LAM = \frac{\sum_{v=v_{\min}}^{N} vP(v)}{\sum_{v=1}^{N} vP(v)} \quad (8)$$

- The average length of vertical lines (TT) is the trapping time and represents the average time in which the system is trapped into a specific state:

$$TT = \frac{\sum_{v=v_{\min}}^{N} vP(v)}{\sum_{v=1}^{N} P(v)} \quad (9)$$

- The length of the longest vertical line (V_{max}) is analogous to L_{max} for the vertical lines:

$$V_{\max} = \left(\{v_i\}_{i=1}^{Nv}\right) \quad (10)$$

From a recurrence plot it is possible to extrapolate the recurrence times [29]. Let us consider the recurrence points of the i^{th} row $\{R_{i,j}\}_{j=1}^{N}$ of an RP which correspond to the set of points of the trajectory which fall into the $\mathcal{E}$ -neighbourhood of an arbitrary chosen point at i. The recurrence times between these recurrence points (*recurrence times of first type*) are:

$$\{T_k^{(1)} = j_{k+1} - j_k\}_{k \in N} \quad (11)$$

Removing all consecutive recurrence points with $T_k^{(1)} = 1$ to avoid tangential motion, the recurrence times of second type are:

$$\{T_k^{(2)} = j'_{k+1} - j'_k\}_{k \in N}, \quad (12)$$

where the set of the remaining recurrence points is used. It turns out that $T^{(2)}$ measures the time distance between the beginning of subsequent recurrence structures in the RP along the vertical direction and it can be considered as an estimate of the average of the lengths of white vertical lines in a column of the plot [19].

A great advantage offered by this analysis is that the calculation of the RQA measures for moving windows along the recurrence plot, allows to identify the transitions of dynamical systems. In particular, it was shown that the positions of the local maxima and local minima in the temporal trends of some measures correspond to chaos-order and chaos-chaos transitions [19].

3. General Framework

The algorithm block scheme is represented in Fig. 1. Since the voice has a non-stationary nature, we perform a short term analysis with a frame size of 40 ms and an overlap of 50 %. Given an input track, an automatic vowel extraction module is used to detect and retain only the vowel frames and for each of them, the optimal parameters (m and τ) for state space reconstruction are found. Then, RPs are generated using the time delay method, and some RQA measures extracted to describe RPs quantitatively. Since a set of RQA measures can be extracted, in principle, for each frame, statistics on these measures may be collected to give a general description of the emotional content of the input sentence.

Each step of the adopted framework is detailed in the following sections.

3.1. Database

The German Berlin Emotional Speech Database (EmoDB) [30] has been employed for all the experiments carried out in this work. The database contains ten sentences pronounced by ten actors (five males and five females) in seven different emotional states: neutral, anger, fear, happiness, sadness, disgust and boredom. The audio tracks were sampled as mono signals at 16 kHz, with 8 bit/sample. Most of the sentences were recorded several times in different versions and the resulting corpus was subjected to a perception test where the degree of recognition of emotions and their naturalness were evaluated by a group of listeners. Utterances with an emotion recognition rate better than 80 % and a naturalness score greater than 60 % were included in the final database. As shown in Table 1, among the 535 available sentences, some emotions prevail over the others. The emotion disgust has been excluded from our analysis because of the too low number of tracks belonging to this group.

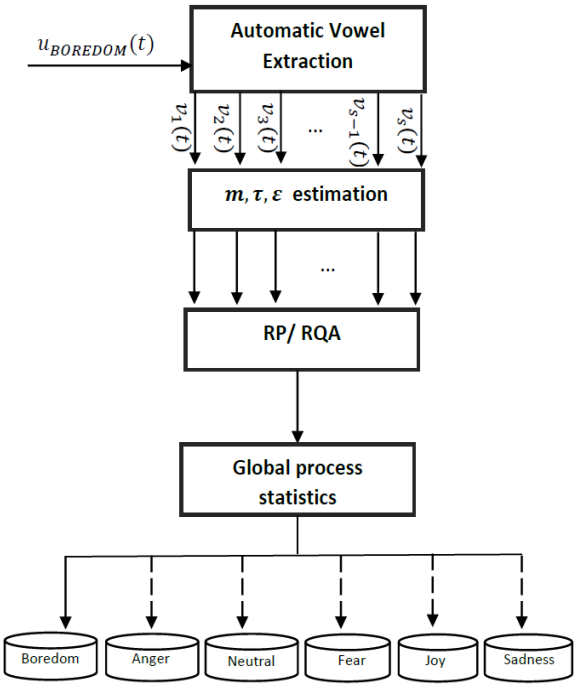

Fig. 1. The algorithm block scheme for an example input sentence.

Table 1. Number of utterances in EmoDB.

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

3.2. Automatic Vowel Extraction

Speech production starts with a compression of the lung volume causing an airflow that is converted in a glottic signal by passing through the vocal folds. This signal is then filtered by the vocal tract and converted into different audible sounds by moving the articulators (i.e. velum, tongue, lips and jaw) [31].

To produce vowel sounds, the vocal tract is open with a uniform cross-sectional area along its length resulting in quasi-periodic sounds. In contrast, when the consonants are pronounced, the vocal tract has a constriction at some point that produces a resistance to the airflow generating turbulent noise. Therefore, the two categories of sounds have very different characteristics that can be highlighted by analyzing the spectral content of the waveforms. The analysis in the frequency domain can be simplified considering that the vocal tract acts as a resonator filter, which has its own resonance frequencies known as formants. By varying the shape of the vocal tract through different combinations of articulations, the formant frequencies of the filter change too. Hence, each vowel sound can be described through its formant frequencies [9].

In particular, it was shown that the distinctive qualities of the vowels can be attributed to differences in the first three formant frequencies and that, very often, the first two formants can univocally identify a vowel [9, 32].

For these reasons, we have extracted some spectral features from the formant frequencies estimated from the power spectral density of the audio track. These features have been used to train a classifier that automatically detects vowel segments in the signal.

Supposing each frame the output of a stationary process, an autoregressive model (AR) has been used to estimate the power spectral density. First, the order of the model has been identified with the Akaike's Information Criterion (AIC) [33] to avoid splitting line and spurious peaks in the final spectrum. Subsequently, the Burg's method [34] has been employed to find the parameters of the AR model. This technique has been preferred over the simple linear prediction analysis as the former identifies the optimal set of parameters by minimizing the sums of squares of the forward and backward prediction errors while the latter uses only the backward errors.

Furthermore, as compared with other parametric methods, the Burg's algorithm ensures more stable models and a higher frequency resolution [35].

The peaks of the power spectral density are in correspondence of the formants position. The first three peaks have been identified in the estimated spectrum and for each of them the following characteristics have been collected:

- The frequency at which they occur;
- The amplitude of the peak;
- The area under the spectral envelope within the -3 dB bandwidth.

To distinguish the vowel sounds from all other types of phonemes (including silence intervals) a one-class classification approach has been adopted. This method was introduced by Schölkopf [36] as a variant of the two-class SVM to identify a set of outliers amongst examples of the single class under consideration. Thus, according to this approach, the outlier data are examples of the negative class (in this case, the not vowels frames). A kernel function is used to map the data into a feature space F in which the origin is the representative point of the negative class. So, the SVM returns a function f that assigns the value +1 in a subspace in which the most of the data points are located and the opposite value -1 elsewhere, in order to separate the examples of the class of interest from the origin of the feature space with the maximum margin.

Formally, let us consider $x_1, x_2, \ldots, x_l$, l training vectors of the one class X, where X is a compact subset of $\mathbb{R}^N$. Let $\Phi : X \to F$ be a kernel function that map the training vectors into another space F. Separating the data set from the origin is equivalent to solving the following quadratic problem:

$$\min_{w \in F, \xi \in \mathbb{R}^l, \rho \in \mathbb{R}} \frac{1}{2}\|w^2\| \frac{1}{vl}\sum_{i=1}^{l}\xi_i - \rho, \quad (13)$$

subject to:

$$\left(w \cdot \Phi(x_i)\right) \geq \rho - \xi_i, \xi_i \geq 0\,, \quad (14)$$

where $v \in (0; 1]$ is a parameter that controls the decision boundary of the classification problem, ξ_i are the nonzero slack variables, w a weight vector and ρ an offset that parametrizes a hyperplane in the feature space associated with the kernel. If w and ρ solve for this problem, then the decision function:

$$f(x) = sign\left(w \cdot \Phi(x) - \rho\right), \quad (15)$$

will be positive for the most of the examples x_i contained in the training set.

Of course, the type of kernel function, the operating parameters of the kernel and the correct value of v, must be estimated to build the one-class SVM classifier. As suggested by the author, we have chosen a Gaussian kernel with Sequential Minimal Optimization (SMO) algorithm to train the classifier, since the data are always separable from the origin in the feature space. For generic patterns x and y, a Gaussian kernel is expressed as:

$$k(x, y) = \exp\left(\frac{\|x - y\|^2}{c}\right), \quad (16)$$

where the parameter c is the kernel scale that controls the tradeoff between the over-fitting and under-fitting loss in the feature space F [37].

Regarding the choice of the value v, it should be taken into account that it represents an upper bound on the fraction of outliers and, at the same time, a lower bound on the fraction of support vectors. It is then necessary to find a value that on the one hand is able to describe the whole dataset for training and on the other hand avoids the over-training of such data. Results on the tuning of the parameters on real data and classification performances are in Section 4.1.

3.3. RP/RQA

The general idea behind all the analysis carried out in this work is that the evolutionary dynamics of each vowel constitute local descriptions of the intrinsic process in the formation of a particular emotion. Therefore, after extraction of vowel segments from a sentence, a frame-level analysis is applied to monitor such dynamics. First, time delays and embedding dimensions are estimated to allow a correct reconstruction of the dynamics in the phase space.

Hence, said s the total number of vowel frames dynamically identified by the Automatic Vowel Extraction module, the time delays vector $\boldsymbol{T} = (\tau_1, \ldots, \tau_s)$ and the embedding dimensions vector $\boldsymbol{M} = (m_1, \ldots, m_s)$ are saved for each sentence. Please note that s is a sentence dependent parameter. At the end, the Recurrence Plots are obtained and the Recurrence Quantitative Analysis is performed on RPs.

In order to explore the time dependent behavior of the recurrence measures, the computation is performed using sliding windows of length W (less than the duration of a frame) with an offset of W_s samples along the main diagonal of the RP of each vowel frame. The values of these two parameters are calculated accounting for the scale of the dynamics to be investigated (local/global) and for the temporal resolution to be achieved [38].

In detail, for the estimation of the window, the smallest value of the first formant among all the vowel frames of the sentence is considered. The choice of W must allow at least the observation of the largest fundamental period:

$$f_{1,min} = \min_{k = 1, \ldots, s}\{f_{1,k}\}$$
$$\widehat{W} = \left\lceil F_c \frac{1}{f_{1,min}} \right\rceil, \quad (17)$$

where F_c is the sampling frequency.

The offset W_s is the embedding window, i.e., the length of the segment of time series that is necessary to reconstruct a single vector in the phase space. Its optimal value has been taken as:

$$W_s = 2 \max_{k = 1, \ldots, s}\{m_k\} \min_{k = 1, \ldots, s}\{\tau_k\}, \quad (18)$$

where τ_k are the elements in $\boldsymbol{T}$, m_k the elements in $\boldsymbol{M}$, while the factor two has been found as a compromise between high time resolution and computational complexity.

The overall trend of each RQA measure is finally reconstructed considering the various vowel segments neatly placed in the sentence (see Fig. 2). For an experimental dataset of sentences, the trends of each RQA measures are grouped by emotion and some statistical tests performed to assess:

1) If the different emotions are statistically different among them and

2) The existence and the nature of relations between the various groups. In addition, some statistics are computed to explore the general characteristics of the emotions expressed in the sentences. The description of the successive analysis is reported in the Section 4.3.

4. Results

The following sections report the performances achieved by the one-class SVM classifier and both qualitative and quantitative results of the recurrence analysis.

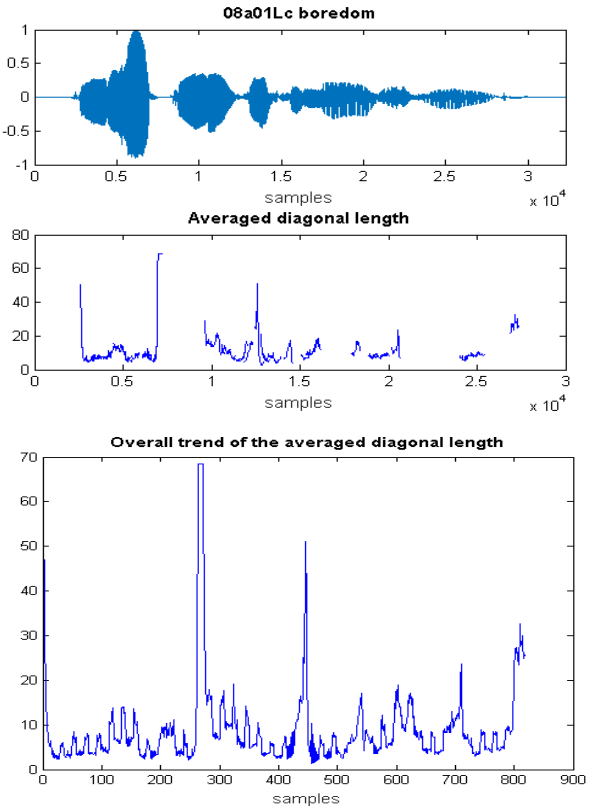

Fig. 2. Example of RQA processing for an input sentence with emotional state boredom. Input track (top); averaged diagonal length computed on the estimated vowels frames (middle); reconstructed trend of the averaged diagonal length (bottom): not-vowel frames and overlapped samples are removed.

4.1. Automatic Vowel Extraction

To train the one class SVM classifier, a dataset was used of 128 segments of German vowels of duration equal to 40 ms, extracted from several sentences spoken by four people (two men and two women) for the six emotions. In order to identify the optimal values for the parameters c and ν, the classifier was trained and validated several times. In particular, due to the nature of the classification problem, an holdout validation scheme has been adopted. So, another set of 83 speech segments including vowels, consonants and pauses, has been used to tune the parameters and identify the most effective model. Keeping fixed the value of ν, the classifier was retrained by varying the value of the kernel scale in a predetermined range. For each model obtained, the performances on the validation set were evaluated in terms of accuracy, sensitivity (or true positive rate), specificity (or true negative rate) and false positive rate. The curves that illustrate the behavior of such measures for three values of ν and by varying the kernel scale from 0 to 2.7 are shown in Fig. 3.

In Fig. 3(b) and Fig 3(c) only one point can be identified to guarantee high performances of the classifier, since the values of accuracy, sensitivity and specificity are high (around 0.7), while the false positive rate remains low. For kernel scale values greater than this optimum, specificity and accuracy decrease rapidly, while sensitivity and false positive rate increase. These results suggest that there is a rapid growth of the number of false positives, i.e., the percentage of the not-vowels frames incorrectly predicted as vowels by the classifier increases.

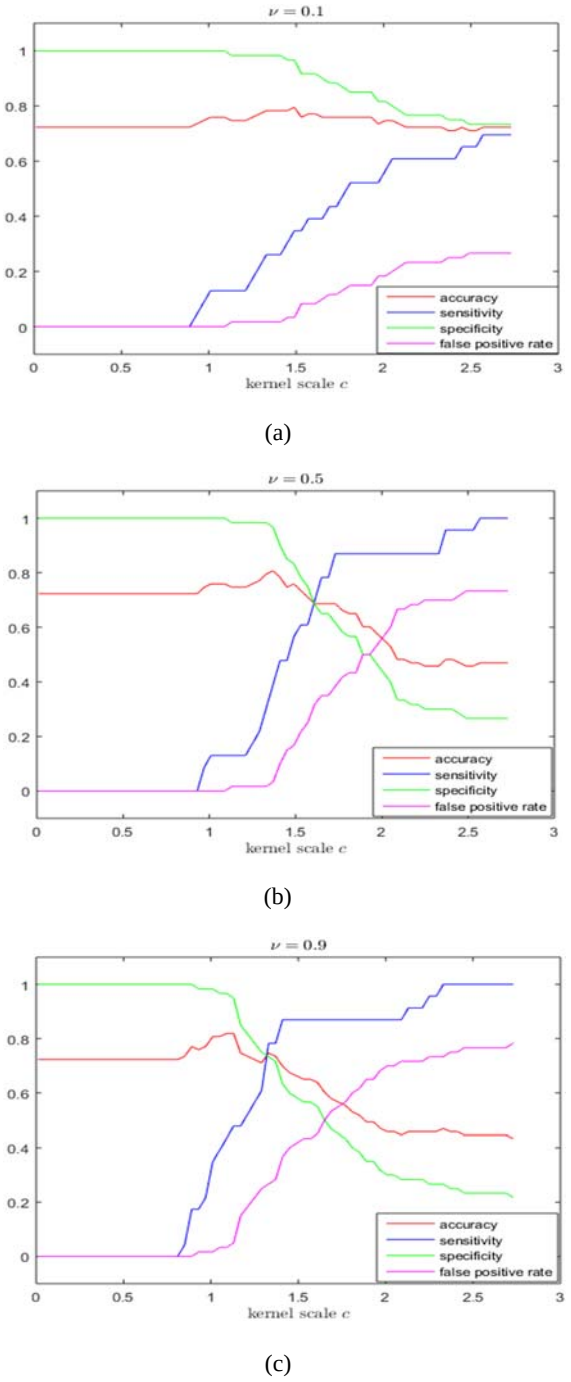

Fig. 3. Accuracy, sensitivity, specificity and false positive rate of the one class SVM classifier in function of the kernel scale c for the fixed parameter (a) $\nu = 0.1$, (b) $\nu = 0.5$ and (c) $\nu = 0.9$.

For our purposes, the system critically depends on the percentage of false positives, since the classifier acts properly if it is capable of rejecting the greatest

amount of not-vowel frames. Therefore, even at the expense of a lower number of true positives and higher percentage of false negatives (vowel frames incorrectly rejected), we have set $\nu = 0.1$ and consequently chosen the value of c at which the classifier returns high values of accuracy and specificity, while maintaining a false positive rate less than 15 % (see Fig. 3(a)).

To assess the performances of the one class SVM with the chosen parameter settings ($\nu = 0.1$ and $c = 1.75$), we performed a final test on a set of 40 speech segments independent of both the training and the validation sets. The confusion matrix is shown in Table 2. As it can be seen, the low rate of false positives (not-vowels incorrectly predicted as vowel frames) confirms the validity of the model for the selected parameters (represented in Fig. 3(a) for $\nu = 0.1$ and $c = 1.75$).

Table 2. Confusion matrix of the one class SVM on the test set composed of 20 vowel and 20 not-vowel frames.

		Predicted conditions	
		Vowels	Not vowels
True Conditions	Vowels	9	11
	Not vowels	4	16

4.2. Qualitative Results: RP

The patterns in RPs can reveal typical behaviors of the system and so they can be used to provide a general description of the time evolution of the dynamic trajectories. Fig. 4 shows the RPs of the vowel /a/ extracted in the same sentence and approximately in the same position, pronounced by a female subject for different emotions. As it can be seen, all RPs have a topology with periodic patterns that are regularly repeated, with the exception of the emotion fear in which there are discontinuities and white bands that indicate the presence of abrupt changes in the dynamics of the system. Another distinctive feature is the length of the diagonal lines: the RPs of boredom and neutral, besides being very similar each other, have the longest diagonal lines; on the other hand, anger and fear show very short diagonal lines. Moreover, a drift can be noted in the emotion sadness: the RP fades away from LOI indicating that the system varies very slowly. The examples show that certain measures are most distinctive for some emotions and that certainly the density of points in the RPs, the length of the lines present in them and measures that are able to differentiate the different kinds of time periodicity (such as T^2), can effectively distinguish among several emotional levels.

4.3. Quantitative Results: RQA

The tracks used to train the one class SVM, together with repeated versions of the same tracks, were excluded from the whole set of tracks in EmoDB to respect the assumption of independent samples required by the statistical tests. The achieved dataset is described in Table 3.

Table 3. Number of tracks in the dataset used for experiments.

Emotion	**No. of utterances**
Anger	82
Boredom	63
Fear	51
Happiness	48
Neutral	62
Sadness	48

In this work, we used the method of false nearest-neighbors to find the embedding dimension m of each vowel frame and the First Local Minimum of Average Mutual Information algorithm to determine the appropriate delay τ. The parameter ε was set to 10 % of the maximum space diameter and a maximum norm was used as norm function. Fig. 5 shows the box plots of the average values of m and τ of the vowels extracted from the sentences analyzed and grouped by emotion. It is interesting to note that, while the average values of m fluctuate around the same value (about 6) for all the emotions and almost in the same way, the distributions of values of τ are different each other, with the exception of boredom, neutral and sadness, which are more similar among them. In the latter case, the box plots suggest that to obtain new and useful information from successive coordinates of the time delayed vectors, it must be considered samples more spaced in the time series of vowels.

The trends of each RQA measure were reconstructed as described in Section 3.3. Then, the trends of all the frames were grouped by emotions for the same RQA, obtaining nine sets of measures (one for each RQA measure. The list of RQA measures is in the first column of Table 4). Each set of measures consists of six groups of data, one for each emotion.
The Shapiro-Wilk test [39] was used to check whether the 54 obtained samples came from normally distributed populations. All the 54 tests returned a $p < 0.0001$ with a significance level $\alpha = 0.05$, so the null hypothesis (normal distribution) was rejected.

Hence, the non-parametric Kruskal-Wallis test [40] was employed as an alternative to one-way ANOVA, for testing whether the six different data groups of each RQA measure originate from the same distribution, at a significance level $\alpha = 0.05$. This test is used in the same way as ANOVA, but it performs for not-normally distributed populations. In order to better appreciate the possible differences among populations, mean, standard deviation (std), median and interquartile range (iqr) values of the nine RQA measures for all the groups of emotions were computed and are reported in Table 4, together with the results of the test χ^2 and the corresponding p-values.

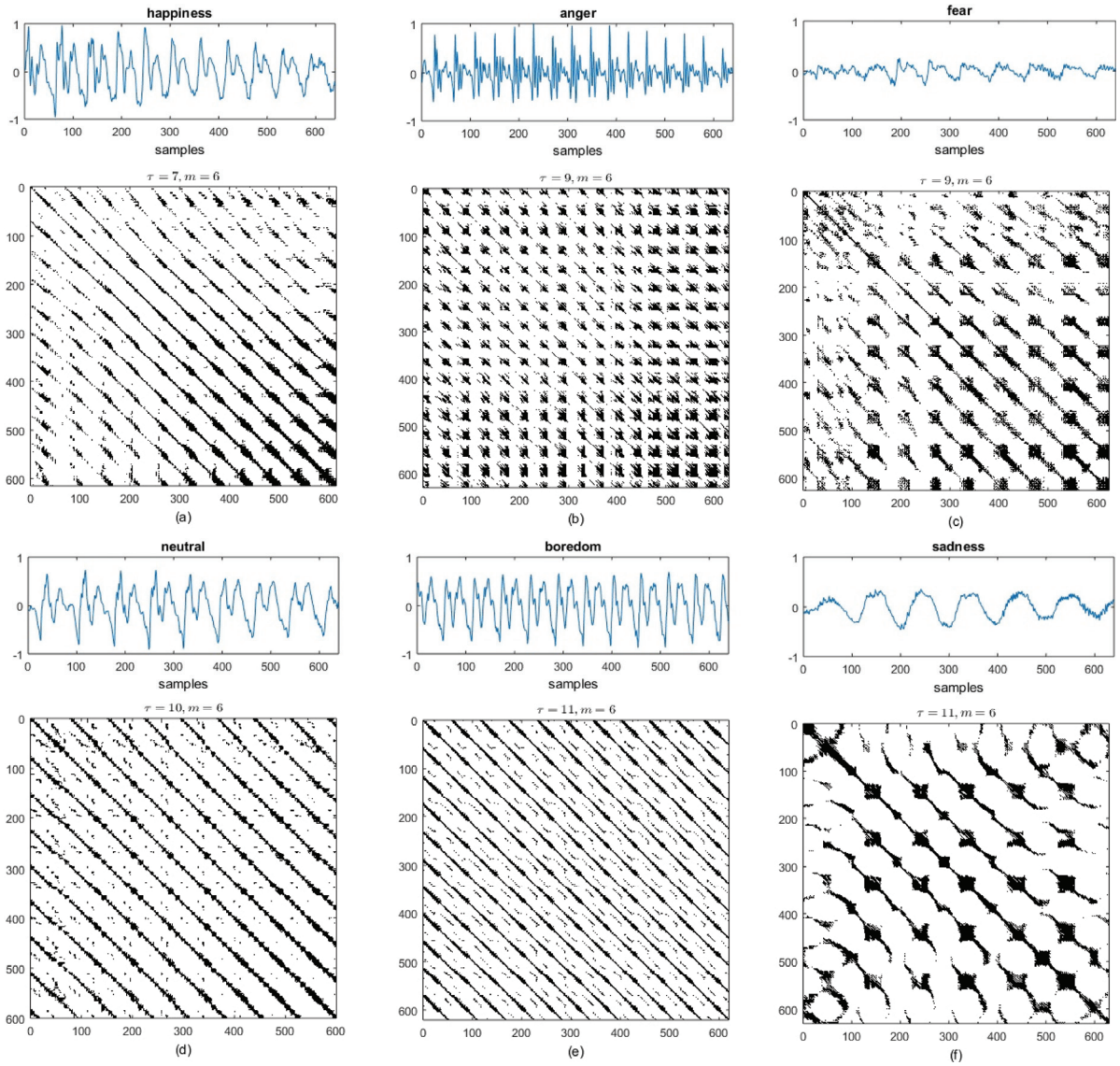

Fig. 4. RPs of vowel /a/ in the track 08a02 for emotions: (a) happiness, (b) anger, (c) fear, (d) neutral, (e) boredom, (f) sadness; ε is setting to 10 % of maximum space diameter with a maximum norm.

A significant Kruskal-Wallis test indicates only that at least one group is statistically different from at least one other group, but it does not identify neither the amount of groups that differ significantly, nor which pairs are statistically different. For the latter purposes, a post-hoc analysis must be performed to compare pairs of groups. Anyway, an inspection of the representative values of the statistics, can give a first impression about the characteristics of the RQA measures.

By considering the medians of the measures with the highest statistics of the test χ^2, it can be stated that:

- for all of them, boredom and neutral exhibit very similar values;
- the measures related to diagonal lines (*RATIO, L, ENTR*) and those concerning the vertical lines (TT, V_{max}), have systematically the following sorting of emotions, in decreasing order of the median values:
  - Sadness;
  - Boredom;
  - Neutral;
  - Fear;
  - Happiness;
  - Anger;
- T^2 is the only measure that show the previous list in the reversed order.

Finally, the Dunn's post-hoc test [41] was chosen to perform multiple pairwise comparisons and determine which groups differ for each measure. The Bonferroni's correction was used to control the family-wise error rate, with a confidence interval of 95 %. The Table 5 provides the p-values of the Dunn's test for all RQA measures. It can be noted that there are significant differences between all emotions, except for boredom-neutral regarding *RATIO, LAM, TT* and V_{max}, and for fear-happiness in case of L ($p > 0.05$). The results achieved from multiple comparisons were confirmed by a permutation test. All the pairs of emotions with stochastic dominance (i.e. $p < 0.05$) were permuted two by two using half

the samples of each group. The permuted samples were randomly selected and the Dunn's post-hoc test was carried out with N = 100 repetitions, finally averaging the p-values of the multiple comparisons test to obtain a single p-value. The results reported a $p > 0.05$, confirming that the two populations are different.

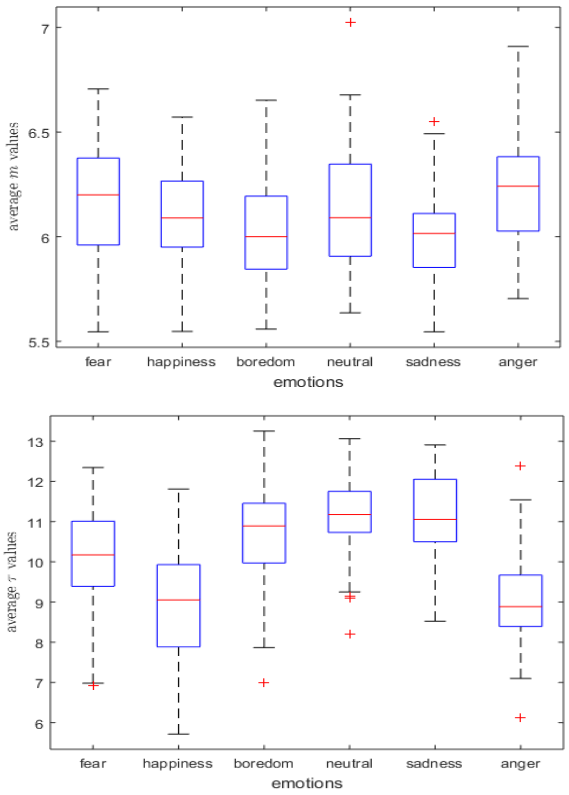

Fig. 5. Average values of m (top) and τ (bottom) of vowel frames grouped by emotion. In each box plot the red line is the position of the median, the edges of the box are the 25th and 75th percentiles, the lower and the upper whiskers represent, respectively, the minimum and the maximum values of the statistics while outliers are plotted individually with red crosses.

5. Discussion and Conclusions

In this work we have investigated the dynamic behavior of vowels taken from a set of spoken sentences of the EmoDB database, for the six emotions anger, boredom, fear, happiness, neutral and sadness.

To extract only the vowel frames, an automatic vowel extraction module was implemented. It consists essentially in a one class SVM classifier that processes the not-vowels frame as outliers. The tuning of the parameters of the classifier and an accurate validation step allowed us to identify a model able to achieve the 79 % of accuracy. We accepted this performance result by considering it as a good compromise between the ability to reject the greatest number of not-vowel frames and that to retain a high number of true positives.

Supposing that the expression of a particular emotional content in a spoken sentence is a gradual complex process, we exploited some properties of the local dynamics of the vowels in it to understand some aspects of the overall process. Firstly, the Takens' embedding theorem was employed to reconstruct the dynamics of each vowel in the phase space. The embedding parameters gave important information about the different emotions: an almost unchanged value of the average of m distributions among the six groups prove that the dimension of the space does not vary and that this parameter can be considered uniform for the general analysis of speech signals. On the other side, the unequal distributions of τ suggests that the trajectories in the phase space are linked to different information rate of the vowels belonging to distinct emotions.

For these reasons, the behavior of the trajectories of vowels dynamics were explored by means of recurrence plots. Different RQA measures were extracted to describe RPs quantitatively. In particular, the computation of these measures was performed by using moving windows along the LOI of the RP of each vowel frame to explore their time dependent behavior. All the reconstructed trends of each RQA measure were grouped by emotion and a statistical analysis was carried out to verify whether these measures were able to describe the different mechanisms underlying the dynamics of each emotion, regardless of the speaker or of the sentence. The multiple pairwise comparisons test has shown that all the RQA measures result statistically significant for discriminating the six groups of emotions with the exception of *RATIO*, *LAM*, *TT* and V_{max} for the couple neutral-boredom and L for the couple fear-happiness.

These results confirm the observations made in Section 4.3, concerning the statistical values shown in Table 4 and the qualitative considerations on the examples shown in Fig. 4: boredom and neutral exhibit both comparable quantitative values and similar graphical patterns. In addition, the RPs of these two emotions are highly diagonal-oriented, so measures based on vertical lines are ineffective for discriminating between the two emotional levels.

In this case, T^2, which is related to the lengths of white vertical lines, can be more efficient. Furthermore, it can be observed that there is a relationship between the rank of the emotions based on the median values of the diagonal and vertical lines-based RQA measures and their levels of activation (or arousal): the emotions with the highest median values are also those with less activation, while emotions with higher activation exhibit lower median values.

In conclusion, it can be observed that certain RQA measures can better discriminate among the basic emotions examined; however it must be hold in consideration that some of them are dependent on each other, so a future development could include a multivariate analysis to identify a subset of measures that perform a better characterization of the different emotional levels. Such measures could be added to the features traditionally used in the literature to try to build a more efficient SER classifier.

Table 4. Means, variances, medians, interquartile range values and Kruskal-Wallis test results. The medians of the RQA measures with the highest statistics of the test χ^2 are shown with bold borders (for a complete discussion see Section 4.3).

		Fear	**Happiness**	**Boredom**	**Neutral**	**Sadness**	**Anger**	**K-W Test**
RATIO	mean	0.13	0.12	0.13	0.14	0.18	0.10	$\chi^2 = 17136.13$ $p = 0$
	std	0.19	0.19	0.17	0.21	0.21	0.18	
	median	0.06	0.05	0.07	0.07	0.10	0.04	
	iqr	0.10	0.09	0.09	0.10	0.13	0.07	
DET	mean	0.65	0.65	0.70	0.71	0.71	0.61	$\chi^2 = 9736.93$ $p = 0$
	std	0.25	0.25	0.24	0.24	0.25	0.25	
	median	0.54	0.55	0.64	0.75	0.63	0.50	
	iqr	0.48	0.47	0.48	0.49	0.49	0.47	
L	mean	8.26	8.07	9.77	10.39	11.73	7.00	$\chi^2 = 16290.25$ $p = 0$
	std	9.48	9.36	9.23	10.25	11.43	8.77	
	median	5.22	5.21	6.86	7.06	8.08	4.51	
	iqr	6.20	5.59	7.53	8.30	9.55	4.65	
L_{max}	mean	57.51	60.69	65.72	63.32	70.23	53.14	$\chi^2 = 6466.76$ $p = 0$
	std	37.16	37.23	35.57	35.86	37.30	37.83	
	median	52.00	54.50	57.50	56.00	64.00	46.50	
	iqr	59.00	61.50	54.00	57.00	51.50	57.00	
ENTR	mean	1.51	1.47	1.67	1.71	1.80	1.33	$\chi^2 = 10993.54$ $p = 0$
	std	0.85	0.85	0.83	0.89	0.93	0.81	
	median	1.28	1.26	1.46	1.49	1.55	1.14	
	iqr	1.09	1.05	1.15	1.21	1.26	0.94	
LAM	mean	0.63	0.61	0.68	0.68	0.69	0.55	$\chi^2 = 10363.09$ $p = 0$
	std	0.26	0.27	0.25	0.25	0.25	0.27	
	median	0.49	0.49	0.49	0.57	0.49	0.49	
	iqr	0.49	0.49	0.49	0.49	0.49	0.48	
TT	mean	4.11	4.84	4.87	5.09	6.28	3.53	$\chi^2 = 18147.99$ $p = 0$
	std	5.11	5.25	5.50	6.21	7.15	5.12	
	median	2.79	2.55	3.35	3.30	4.08	2.35	
	iqr	2.56	2.31	3.12	3.21	4.42	2.06	
V_{max}	mean	11.90	10.87	14.13	14.96	19.00	10.25	$\chi^2 = 19165.64$ $p = 0$
	std	16.64	17.40	17.62	19.28	22.03	17.31	
	median	7.00	6.00	8.00	8.00	11.00	5.00	
	iqr	8.00	6.50	10.00	11.00	16.00	7.00	
$T^{(2)}$	mean	15.78	18.69	12.31	12.06	11.12	19.96	$= 31404.79$ $p = 0$
	std	10.43	10.53	9.81	9.89	9.92	11.22	
	median	13.83	16.26	9.34	9.08	7.78	17.27	
	iqr	14.99	14.92	11.90	11.93	10.19	15.20	

Table 5. p – values of the Dunn's multiple comparison test for all RQA measures.

Couple of emotions	*RATIO*	*DET*	*L*	L_{max}	*ENTR*	*LAM*	*TT*	V_{max}	$T^{(2)}$
Anger - Boredom	*	*	*	*	*	*	*	*	*
Anger - Fear	*	*	*	*	*	*	*	*	*
Anger - Happiness	*	*	*	*	*	*	*	*	*
Anger - Neutral	*	*	0.0006	*	*	*	*	*	*
Anger - Sadness	*	*	*	*	*	*	*	*	*
Boredom - Fear	*	*	*	*	*	*	*	*	*
Boredom- Happiness	*	*	*	*	*	*	*	*	*
Boredom - Neutral	1.00	*	*	*	0.0001	1.00	0.07	1.00	*
Boredom - Sadness	*	*	*	*	*	*	*	*	*
Fear - Happiness	*	*	*	*	*	*	*	*	*
Fear - Neutral	*	*	*	*	*	*	*	*	*
Fear - Sadness	*	*	*	*	*	*	*	*	*
Happiness - Neutral	*	*	*	*	*	*	*	*	*
Happiness - Sadness	*	*	*	*	*	*	*	*	*
Neutral - Sadness	*	*	*	*	*	*	*	*	*

* Statistically significant with a $p - value < 2 \cdot 10^{-16}$

References

[1]. R. W. Picard, Affective Computing, *MIT Press*, Cambridge, 1997.

[2]. K. R. Scherer, Vocal affect expression: A review and a model for future research, *Psychological Bulletin*, Vol. 99, Issue 2, 1986, pp. 143-165.

[3]. C. N. Anagnostopoulos, T. Iliou, I. Giannoukos, Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011, *Artificial Intelligence Review*, Vol. 43, Issue 2, 2015, pp. 155-177.

[4]. T. Vogt, E. André, J. Wagner, Automatic Recognition of Emotions from Speech: A Review of the Literature and Recommendations for Practical Realization, Affect and Emotion in Human-Computer Interaction, *Lecture Notes in Computer Science*, Vol. 4868, 2008, pp. 75-91.

[5]. M. E. Ayadi, M. S. Kamel, F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, *Pattern Recognition*, Vol. 44, Issue 3, 2011, pp. 572-587.

[6]. I. Luengo, E. Navas, I. Hernáez, Feature Analysis and Evaluation for Automatic Emotion Identification in Speech, *IEEE Transactions on Multimedia*, Vol. 12, Issue 6, 2010, pp. 490-501.

[7]. M. Lugger, B. Yang, The relevance of voice quality features in speaker independent emotion recognition, in *Proceedings of the International Conference on Acoustics, Speech and Signal Processing*, Honolulu, HI, Vol. 4, April. 2007, pp. 17-20.

[8]. J. L. Flanagan, Source-system interaction in the vocal tract, *Annals of the New York Academy of Sciences*, Vol. 155, Issue 1, 1968, pp. 9-17.

[9]. G. Fant, Acoustic Theory of Speech Production, *Mouton & Co*., The Hague, 1960.

[10]. I. R. Titze, R. Baken, H. Herzel, Evidence of chaos in vocal fold vibration, I. R. Titze. Ed., Vocal Fold Physiology: New Frontiers in Basic Science, *Singular Publishing Group*, San Diego, 1993, pp. 88-143.

[11]. H. Herzel, D. Berry, I. Titze, I. Steinecke, Nonlinear dynamics of the voice - signal analysis and biomechanical modeling, *Chaos*, Vol. 5, 1995, pp. 30-34.

[12]. A. Giovanni, M. Ouaknine, R. Guelfucci, T. Yu, M. Zanaret, J. M. Triglia, Nonlinear behavior of vocal fold vibration: the role of coupling between the vocal fold, *Journal of Voice*, Vol. 13, Issue 4, 1999, pp. 465-476.

[13]. H. Herzel, Bifurcations and Chaos in Voice Signals, *Applied Mechanics Reviews*, Vol. 46, Issue 7, 1993, pp. 399-413.

[14]. P. Henriquez, *et al*., Characterization of Healthy and Pathological Voice Through Measures Based on Nonlinear Dynamics, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 17, Issue 6, 2009, pp. 1186-1195.

[15]. J. D. Arias-Londoño, J. I. Godino-Llorente, N. Sáenz-Lechón, V. Osma-Ruiz, G. Castellanos-Domínguez, Automatic Detection of Pathological Voices Using Complexity Measures, Noise Parameters, and Mel-Cepstral Coefficients, *IEEE Transactions on Biomedical Engineering*, Vol. 58, Issue 2, 2011, pp. 370-379.

[16]. H. Poincaré, Sur la probleme des trois corps et les équations de la dynamique, *Acta Mathematica*, Vol. 13, 1890, pp. 1-271.

[17]. J. P. Eckmann, S. O. Kamphorst, D. Ruelle, Recurrence plots of dynamical systems, *Europhysics Letters*, Vol. 5, No. 9, 1987, pp. 973-977.

[18]. J. P. Zbilut, C. L. Webber Jr., Embeddings and delays as derived from quantification of recurrence plots, *Physics Letters A*, Vol. 171, Issues 3-4, 1992, pp. 199-203.

[19]. N. Marwan, M. C. Romano, M. Thiel, J. Kurths, Recurrence Plots for the Analysis of Complex Systems, *Physics Reports*, Vol. 438, Issues 5-6, 2007, pp. 237-329.

[20]. C. L. Webber Jr., N. Marwan, Recurrence Quantification Analysis - Theory and Best Practices, *Springer*, Cham, 2015.

[21]. J. P. Bello, Measuring Structural Similarity in Music, *IEEE Transactions on Audio, Speech, and Language Processing*, Vol. 19, Issue 7, 2011, pp. 2013-2025.

[22]. M. A. Little, P. E. McSharry, S. J. Roberts, D. A. Costello, I. M. Moroz, Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection, *BioMedical Engineering OnLine*, Vol. 6, Issue 1, 2007.

[23]. A. Lombardi, P. Guccione, Analysis of emotions in vowels: a recurrence approach, in *Proceedings of the First International Conference on Advances in Signal, Image and Video Processing (SIGNAL'16)*, Lisbon, Portugal, 25-30 June 2016, pp. 27-32.

[24]. G. L. Baker, J. P. Gollub, Chaotic Dynamics: An Introduction, *Cambridge University Press*, Cambridge, 1990.

[25]. F. Takens, Detecting strange attractors in turbulence, *Lectures Notes in Mathematics*, Vol. 898, 1981, pp. 366-381.

[26]. A. M. Fraser, H. L. Swinney, Independent coordinates for strange attractors from mutual information, *Physical Review A*, Vol. 33, Issue 2, 1986, pp. 1134-1140.

[27]. M. B. Kennel, R. Brown, H. D. I. Abarbanel, Determining embedding dimension for phase-space reconstruction using a geometrical construction, *Physical Review A*, Vol. 45, Issue 6, 1992, pp. 3403-3411.

[28]. G. M. Mindlin, R. Gilmore, Topological analysis and synthesis of chaotic time series, *Physica D*, Vol. 58, Issues 1-4, 1992, pp. 229-242.

[29]. J. B. Gao, Recurrence time statistics for chaotic systems and their applications, *Physical Review Letters*, Vol. 83, Issue 16, 1999, pp. 3178-3181.

[30]. F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A database of German emotional speech, in *Proceedings of the Conference on Interspeech*, Lisbon, Portugal, 2005, pp. 1517-1520.

[31]. J. E. Clark, C. Yallop, An Introduction to Phonetics and Phonology, *Wiley-Blackwell*, Oxford, 1995.

[32]. J. C. Catford, A Practical Introduction to Phonetics, *Oxford University Press*, 1988.

[33]. H. Akaike, A new look at the statistical model identification, *IEEE Transaction on Automatic Control*, Vol. 19, Issue 6, 1974, pp. 716-723.

[34]. J. P. Burg, Maximum entropy spectral analysis, in *Proceedings of the 37th Meeting of Society of Exploration Geophysicists*, Oklahoma City, 1967.

[35]. B. I. Helme, Ch. L. Nikias, Improved spectrum performance via a data-adaptive weighted Burg technique, *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 33, Issue 4, 1985, pp. 903-910.

[36]. B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, R. C. Williamson, Estimating the support of a high-

dimensional distribution, Technical report MSR-TR-99-87, *Microsoft Research Corporation*, 1999, pp. 1-28.
[37]. Q. Chang, Q. Chen, X. Wang, Scaling Gaussian RBF kernel width to improve SVM classification, in *Proceedings of the International Conference on Neural Networks and Brain*, Beijing, China, Vol. 1, October 2005, pp. 19-22.
[38]. C. L. Webber Jr., J. P. Zbilut, Recurrence quantification analysis of nonlinear dynamical systems, in Tutorials in contemporary nonlinear methods for the behavioral sciences, M. A. Riley & G. C. Van Orden (Eds.), *National Science Foundation*, 2005, pp. 26-94.
[39]. S. S. Shapiro, M. B. Wilk, An analysis of variance test for normality (complete samples), *Biometrika*, Vol. 52, Issues 3-4, 1965, pp. 591-611.
[40]. W. Kruskal, W. A. Wallis, Use of ranks in one-criterion variance analysis, *Journal of the American Statistical Association*, Vol. 47, Issue 260, 1952 pp. 583-621.
[41]. O. J. Dunn, Multiple comparisons using rank sums, *Technometrics*, Vol. 6, No. 3, 1964, pp. 241-252.