DETECTION OF EMOTIONS AND STRESS THROUGH SPEECH ANALYSIS

Inmaculada Mohino
inmaculada.mohino@edu.uah.es

Maria Goñi
m.goni@edu.uah.es

Lorena Álvarez
lorena.alvrezp@uah.es

Cosme Llerena
cosme.llerena@uah.es

Roberto Gil Pita
roberto.gil@uah.es

University of Alcalá
Signal Theory and Communications Department
28805 Alcalá de Henares (Madrid)
SPAIN

ABSTRACT
The term “stress” is referred to a psychological state, obtained in response to a situation which is accompanied by an emotional reaction, such as, for instance, anxiety, anger, sadness, etc. Emotions are psychophysiological reactions representing modes of adaptation to certain either environmental stimuli or own stimuli. They are those feelings or perceptions of the elements and relations of the reality and the imagination, which are physically expressed by some physiological functions, like, for example, facial reactions, changes in heartbeat or distortions in the paralinguistic aspects of speech. Although both stress and emotional state do not alter the linguistic content, they could affect the paralinguistic contents of speech and this is an important factor in human communication, because it provides more information about the interlocutor than the merely semantic one.

This subliminal information of speech is analyzed in this paper. Throughout the present document many features and algorithms are studied, which are combined to obtain an emotion and stress detector with a high degree of reliability.

KEY WORDS
Stress/emotion detection, signal processing, speech.

1 Introduction

Emotional speech recognition aims at automatically identifying the emotional or physical state of a human being from his or her voice. Just in this respect, early research on the effects of emotions on behavior and language was briefly described by Charles Darwin in his book “The Expression of the Emotions in Man and Animals” [1]. The emotional and physical states of a speaker are known as emotional aspects of speech and are included in the so-called paralinguistic aspects. Although the emotional state does not alter the linguistic content, it is an important factor in human communication, because it provides feedback information in many applications as the one proposed in this paper.

In the literature, it has been widely explored the idea of designing a classification approach that aims to recognize emotions from speech [2]. In this sense, it has been also studied the relationship between voice parameters and emotions. The basic idea underlying these studies is to identify some parameters of the voice that help the classification system to identify emotions. Some of the parameters that have been identified as appropriate features to express emotions are, among others, pitch or fundamental frequency, duration or some related voice quality parameters. Currently, research is focused on finding robust combinations of features and/or classifiers that advance the classification efficiency in real life applications.

The question underlying here is what is the relationship between emotions and stress? Addressing this question demands to mention that emotions and stress are closely related. Although it will be better understood later on, we can say in advance that emotions may be categorized as different levels of stress. It seems to be clear that some emotions like anger or fear result in a level of stress higher than other emotions like, for example, sadness or boredom.

With this in mind, this paper explores the use of a classification approach for the purpose of automatically classifying the input speech signal into one emotion or another (a total of seven emotions have been considered: anger, happiness, disgust, anxiety/fear, sadness, boredom, neutral), and consequently, into a different level of stress. To achieve this, not only do we propose a novel feature that improves the percentage of correct classification but also we study a variety of different emotion classification tasks in order to enhance the overall performance of the classification system. As will be shown, since there is a large number of speech signal-describing features which feed the automatic classifier, it is also explored here the use of a genetic algorithm (GA) for the selection of the most appropriate ones.

The remainder of this paper is structured as follows. First, Section 2.1 will introduce the implemented classification system, describing the feature extraction block (Subsection 2.1) and the classifier evaluated in this paper (Subsection 2.2). Section 3 will describe the implementation of
the GA that performs the feature selection. Section 4 will show the results obtained and finally, the paper will conclude with a brief conclusion on the results.

2 The Proposed System

It basically consists of a feature extraction block, and the aforementioned emotion classifier.

2.1 Feature extraction

The feature extraction task plays the key role of processing the input speech signal in order to extract some kind of valuable information that helps the classifying algorithm to properly discriminate the input signal among the emotions considered. In this respect, there is a number of interesting features that could potentially exhibit different behavior for the aforementioned emotions. These features will be now briefly described.

- **Mel-Frequency Cepstral Coefficients (MFCCs):**
  The MFCCs are a set of perceptual parameters calculated from the STFT [4] that have been widely used in speech recognition. They provide a compact representation of the spectral envelope, so that most of the signal energy is concentrated in the first coefficients. Perceptual analysis emulates human ear non-linear frequency response by creating a set of filters on non-linearly spaced frequency bands. Mel cepstral analysis uses the Mel scale and a cepstral smoothing in order to get the final smoothed spectrum.

  To represent speech, 13 coefficients are commonly used, although it has been demonstrated that for classification tasks, it is enough to take into account only the first 5 coefficients [3] and this is basically the reason why we have chosen 5 MFCCs in this work.

- **Delta Mel-Frequency Cepstral Coefficients (ΔMFCCs):**
  These simple features are calculated from MFCCs. From a mathematical point of view, they are computed as follows:

  \[ \Delta MFCC_n = MFCC_{n-2} - MFCC_{n+2} \]  

  where \( n = 1, \ldots, 5 \) is the index over the 5 MFCCs previously computed. Please note that if there is not exist an MFCC, it is set to be 0. In the same line of reasoning as that in MFCCs, we have computed 5 ΔMFCCs.

- **Short therm energy:** The short term energy is calculated by the evaluation of the energy of the signal within each analysis frame.

- **Pitch:** This measurement, also known as fundamental frequency (F0), is related to the vibration of the vocal cords [5]. In order to estimate the pitch of each analysis frame, we calculate 10 linear prediction coefficients. By filtering the analysis frame with these linear prediction coefficients, the prediction error may be obtained. Then, the autocorrelation of this prediction error is evaluated. The pitch causes a periodicity peak in the autocorrelation, which can be analyzed. If this peak is higher than 20% of the maximum value of the autocorrelation, the frame is considered as voiced, and the value of the pitch is extracted. Otherwise, the frame is considered as unvoiced and consequently, the pitch is not computed.

- **Pitch Perturbation Quotient (PPQ):** This parameter calculates the relative variability period to period of the fundamental frequency within the input speech signal with a smoothing factor of \( m \) frames. Specifically, it averages the differences between the values of the study-case frame to those of \((m-1)/2 \) frames and \((m-1)/2 \) successive frames [6]. Then, the same calculation is repeated for all sets of \( m \) adjacent frames using a moving window. From a mathematical point of view, it can be computed as shown in Equation (2):

  \[
  PPQ_m(\%) = \frac{L - \frac{m-1}{2} \sum_{i=\frac{m+1}{2}}^{\frac{m+1}{2}} \left| F_i - \frac{1}{m} \sum_{k=\frac{m-1}{2}}^{\frac{m-1}{2}} F_k \right|}{(L - (m - 1)) \frac{1}{L} \sum_{i=1}^{L} F_i} \cdot 100
  \]  

  \( F_i, i = 1, \ldots, L \) being the pitch for each frame \( i \)-th. In this work, we have used \( m = 5 \) because it has been empirically observed that this value gives good results.

- **Amplitude Perturbation Quotient (APQ):** This measure is a relative evaluation of the period-to-period variability of the peak-to-peak amplitude within the analyzed frame at smoothing of \( m \) frames. We shall mathematically define this measure as depicted in Equation (3):

  \[
  APQ_m(\%) = \frac{L - \frac{m-1}{2} \sum_{i=\frac{m+1}{2}}^{\frac{m+1}{2}} \left| E_i - \frac{1}{m} \sum_{k=\frac{m-1}{2}}^{\frac{m-1}{2}} E_k \right|}{(L - (m - 1)) \frac{1}{L} \sum_{i=1}^{L} E_i} \cdot 100
  \]  

  Please note that we have used energy values \( (E) \) instead of amplitude. We have also used \( m = 5 \) because it has been empirically observed that this value gives good results.
• **Jitter:** The absolute jitter [8] is the period-to-period variation of the fundamental frequency, i.e. the mean absolute difference (MAJ) between two consecutive pitches. From a mathematical point of view, it can be calculated as stated in Expression (4)

\[
MAJ = \frac{1}{L-1} \sum_{i=1}^{L-1} |F_i - F_{i+1}|
\]  

(4)

where \(F\) are the extracted pitches and \(L\) is the number of frames.

In this work, we have made use of the relative jitter, which is defined as the mean absolute difference between consecutive pitches divided into the average pitch, as illustrated in Equation (5):

\[
JITTER(\%) = \frac{MAJ}{F_{av}} \cdot 100.
\]

(5)

• **Shimmer:** The absolute shimmer [8] is a measure of period-to-period variability of the energy value, expressed as written in Equation (6)

\[
MAS = \frac{1}{L-1} \sum_{i=1}^{L-1} |E_i - E_{i+1}|
\]

(6)

where \(E\) are the extracted energies and \(L\) being the number of frames.

In this work, we have made use of the relative shimmer which is defined as the mean absolute difference between consecutive energies, divided by the average energy. It is expressed as a percentage, as illustrated in Equation (7)

\[
SHIMMER(\%) = \frac{MAS}{E_{av}} \cdot 100
\]

(7)

• **Harmonic Noise Rate (HNR):** One of the important characteristics of voiced speech is the well-defined harmonic structure. The source for the voiced speech is often modeled as quasi-periodic glottal pulses. But in reality, even the sustained vowel phonation consists of some random parts mainly due to turbulence of airflow through the glottis and due to pitch perturbations. HNR quantifies the degree of that noisiness by comparing the amount of energy produced in the vocal cords versus the non-harmonic energy contained in the signal. A variety of such measures have also been developed, with Boersma’s harmonicity measure [9] arguably being the most successful, which works under the assumption that the signal to be analyzed is periodic, which in the case of speech segments very short of the voice signal may be considered practically certain. In the following paragraphs, it is explained in a detail the way the HNR is computed in this paper.

Let us assume that \(r(\tau)\) is the autocorrelation of the signal \(x(t)\), then \(r(0)\) represents the total energy of the signal with periodic components \((r_p)\) and aperiodic components \((r_{ap})\).

\[
r(\tau) = r_p + r_{ap}.
\]

(8)

With this in mind, the normalized autocorrelation, \((r_x)\) is calculated as shown in Equation (9)

\[
r_x(\tau) = \frac{r(\tau)}{r(0)}
\]

(9)

So, given the frequency of the periodic components of the autocorrelation and assuming additive white noise and the voice, the energy of the periodic components, \(r_{xp}\) can be calculated as follows:

\[
r_{xp} = r_x(T_0)
\]

(10)

Or in other words, it represents the value of the normalized autocorrelation in \(n = T_0\). Since the total power is set to be 1, then the noise energy (or in other words, the energy due to aperiodic components) corresponds with the difference between the total energy and the corresponding to the periodic components:

\[
r_{xap} = 1 - r_{xp}.
\]

(11)

With the aforementioned components in mind, we finally obtain Expression (12):

\[
\text{hnr} = \frac{r_{xp}}{1 - r_{xp}} = \frac{r_x(T_0)}{1 - r_x(T_0)} = \frac{r(T_0)}{r(0) - r(T_0)}
\]

(12)

where the HNR (expressed in dB) is obtained by using the expression shown below:

\[
\text{HNR (dB)} = 10 \log_{10}(\text{hnr}).
\]

(13)

• **Voice Turbulence Index (VTI):** Although there are different definitions of the frequency margins, VTI is commonly defined as the ratio of the spectral inharmonic high-frequency energy ranging from 2800 Hz to 5800 Hz to the spectral harmonic energy in the range of 70 to 500 Hz [10].

In the experimental work carried out in this work, the use of the frequency ranges 50-2500 Hz (for spectral harmonic energy) and 2500-5800 Hz (for spectral inharmonic high-frequency energy) has been proven to be very convenient to improve classification results.

In order to calculate VTI, it is necessary to previously compute the following two parameters:

- LFEx: it labels the spectral energy from frequency bands corresponding to 50-2500 Hz.
- HFEx: it labels the spectral energy from frequency bands corresponding to 2500-5800 Hz.

With these two parameters in mind, the spectral harmonic energy is computed as follows:

\[
HLEF_x = LFE_x \cdot r_x(T_0) \quad (14)
\]

and VTI is defined as follows:

\[
VTI = \frac{HLEF_x}{HFE_x} = \frac{HFE_x}{LFE_x \cdot r_x(T_0)} \quad (15)
\]

**Soft Phonation Index (SPI):**

This parameter is not a measurement of noise but rather the harmonic structure of the spectrum. SPI is an average ratio of the lower spectral harmonic energy (70-1600 Hz) to the higher frequency (1600-4500 Hz) harmonic energy [10].

To compute SPI, it is necessary to define:

- \( r_H(T_0) \): Normalized autocorrelation value in \( T_0 \) of the high frequency signal.
- \( r_L(T_0) \): Normalized autocorrelation value in \( T_0 \) of the low frequency signal.

By means of the two aforementioned parameters, SPI is computed as shown below:

\[
SPI = \frac{r_L(T_0)}{r_H(T_0)} \quad (16)
\]

**DATA:** This parameter simply refers to the input digitized speech signal.

**Energy Measurement versus Ratio (EMR):** This measure is computed as stated in Equation (17)

\[
EMR = \frac{E_x}{R} \quad (17)
\]

where \( E_x \) labels the energy of the input speech signal and \( R \) is the range of the energy defined as \( R = E_{max} - E_{min} \).

**Normalized Harmonic Energy Variation (NHEV):**

This novel measurement is determined from a spectral analysis of the variations of \( r_x(T_0) \). First, we evaluate the Discrete Fourier Transform (DFT) of the sequence of terms \( r_x(T_0) \) obtained for each time frame. So, we have a spectral representation of the time variations of \( r_x(T_0) \). Using this spectrum, we are able to determine \( EH_{Har}F_L \), the energy of the frequency components ranging from 0.1Hz to 4Hz. The NHEV is then obtained using equation (18), where \( LFR \) is the energy the original signal from 2.5 kHz to 5.8 kHz for each frame.

This is a proposed measurement. Harmonic energy signal \( r_x(T_0) \) is calculated as was previously explained and the energy from 0.1Hz to 4Hz is determined using the Discrete Fourier Transform (DFT), being named like \( EH_{Har}F_L \). On the other hand, the harmonic energy the original signal from 2.5 kHz to 5.8 kHz for frame is calculated, which name is \( LFR \).

Finally it is obtained:

\[
NHEV = \frac{LFR}{EH_{Har}F_L} \quad (18)
\]

Since the input speech signal has been divided into a number of frames of \( T = 20 \) ms, the aforementioned measurements, except PPQ, APQ and EMR, have been calculated for any frame. The statistical measures considered in this work include mean, variance, standard deviation, kurtosis, skewness, the logarithmic mean and the geometric mean. Therefore, the feature extraction algorithm generates the following feature vector: \( F = [sp[M_1], sp[M_1], sp[M_2], sp[M_2], ..., sp[M_T], sp[M_T]] \), where \( sp \) labels a statistical parameter and \( T \) the total number of measurements considered. This is just the signal-describing vector that feeds the classifier, its dimension being \( dim(F) = 57 \). For the sake of clarity, it is written formally \( F = [F_1, ..., F_{57}] \). For properly completing this section, Table 1 summarizes the features considered in this work. Please note that not all the statistical parameters are estimated for all the measurements.

2.2 Mean Square Error (MSE) Linear Classifier

Linear classifiers are characterized by the use of linear decision boundaries, which implies that they cannot discriminate among classes associated in very complex shapes. In this paper, we observed that the results are very satisfactory for the particular classification task at hand.

Let us consider a set of training patterns \( \mathbf{x} = [x_1, x_2, ..., x_L]^T \), where each of these patterns is assigned to one of the possible classes denoted as \( C_i, i = 1, ..., K \). In a linear classifier, the decision rule is obtained using a set of \( K \) linear combinations of the training patterns, as it can be observed in Equation (19).

\[
y_k = w_{k0} + \sum_{n=1}^{L} w_{kn}x_n \quad (19)
\]

Where \( w_{kn} \) are the weighting values and \( w_{k0} \) is the threshold. Furthermore, Equation (19) can be expressed in matrix notation as shown in Equation (20).

\[
\mathbf{y} = \mathbf{w}_0 + \mathbf{W}^T \mathbf{x} \quad (20)
\]

Where \( \mathbf{W} \) is the weight matrix that contain the values of \( w_{kn} \). The design of the classifier consists of finding the best values of \( \mathbf{W} \) and \( w_0 \) in order to minimize the classification error.
Table 1. A summary of the features used in this work.

<table>
<thead>
<tr>
<th>Features</th>
<th>Statistical parameters</th>
<th>Total number</th>
</tr>
</thead>
<tbody>
<tr>
<td>MFCC</td>
<td>mean (·) std (·)</td>
<td>10</td>
</tr>
<tr>
<td>ΔMFCC</td>
<td>mean (energy) std (energy) mean (pitch) std (pitch) mean (pitch<em>energy) std (pitch</em>energy) mean (pitch<em>energy) std (pitch</em>energy) mean (pitch<em>energy) std (pitch</em>energy)</td>
<td>5</td>
</tr>
<tr>
<td>PITCH and ENERGY</td>
<td>mean (·) std (·)</td>
<td>10</td>
</tr>
<tr>
<td>SHIMMER</td>
<td>mean (·) mean (·)</td>
<td>2</td>
</tr>
<tr>
<td>HNR</td>
<td>mean (·) std (·) geomean (·) var (·) kurtosis (·) skewness (·)</td>
<td>6</td>
</tr>
<tr>
<td>VTI</td>
<td>mean (·) mean (·)</td>
<td>4</td>
</tr>
<tr>
<td>SPI</td>
<td>mean (·) geomean (·) mean (log(·))</td>
<td>3</td>
</tr>
<tr>
<td>DATA</td>
<td>var (·) std (·) kurtosis (·) skewness (·) mean (·)</td>
<td>5</td>
</tr>
<tr>
<td>EMR</td>
<td>mean (·) std (·) var (·) geomean (·) mean (log(·))</td>
<td>1</td>
</tr>
<tr>
<td>NHEV (Proposed feature)</td>
<td>mean (·) std (·) var (·) geomean (·) mean (log(·))</td>
<td>7</td>
</tr>
</tbody>
</table>

The output of the linear combinations $y$ is used to determine the decision rule. For instance, if the component $y_k$ gives the maximum value of the vector, then the $k$-th class is assigned to the pattern.

In order to determine the values of the weights, it is necessary to minimize the mean squared error value. Let us define the matrix $V = [w_0, W]^T$ containing the weight matrix $W$ and the threshold vector $w_0$, then the pattern matrix $Q$, which contains the input features for classification, is expressed in Equation (21).

$$Q = \begin{bmatrix} 1 & 1 & 1 & \ldots & 1 \\ x_{11} & x_{12} & x_{13} & \ldots & x_{1N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{L1} & x_{L2} & x_{L3} & \ldots & x_{LN} \end{bmatrix}$$ (21)

So, the output of the linear classifier is obtained as a linear combination of the inputs according to Expression (22).

$$Y = V \cdot Q$$ (22)

Let us now define the target matrix containing the labels of each pattern as:

$$T = \begin{bmatrix} t_{11} & t_{12} & t_{13} & \ldots & t_{1N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_{K1} & t_{K2} & t_{K3} & \ldots & t_{KN} \end{bmatrix}$$ (23)

where $N$ is the number of data samples, and $t_{kn} = 1$ if the $n$-th pattern belongs to class $C_k$, and 0 in other case.

Then, the error is the difference between the outputs of the classifier and the correct values, which are contained in the target vector:

$$E = Y - T = V \cdot Q - T$$ (24)

Consequently, the mean square error (MSE) is computed according to Equation (25).

$$MSE = \frac{1}{N} \|Y - T\|^2 = \frac{1}{N} \|V \cdot Q - T\|^2$$ (25)

In the least squares approach, the weights are adjusted in order to minimize the mean square error. The minimization of the MSE is obtained deriving expression (25) with respect $V$ and, using the equations of Wiener-Hopf [11], the next expression for the weight values is obtained:

$$V = T \cdot Q^T \cdot \left( Q \cdot Q^T \right)^{-1}.$$ (26)

This expression allows to determine the values of the coefficients that minimize the mean square error for a given set of features.
3 Selecting Feature by Using Genetic Algorithms

As was explained in-depth in [12], a GA is an optimization and search technique which, inspired by the principles of genetics and natural selection, exhibits useful properties for solving certain problems that otherwise would be intractable, such as, for instance the one proposed here.

What is the analogy to our problem? In the problem at hand, we are looking for the best feature vector that helps the system to perform a proper classification. In this search, a group (or “population”, using genetic information) are evaluated with the aim of maximizing the percentage of correct classification. How well the trial solution solves the problem, i.e., how accurate the system classifies, is the so-called “fitness of the individual”.

For this particular terminology to be understood, it is important to have a brief look at the way any trial feature vector is codified. In nature, all the genetic information which encodes and causes the external characteristics of a living organism (or individual) is called genotype. Any particular characteristic produced by a piece of this genetic information is encoded by a “gene”, the “chromosome” being the set of these genes. Each gene is located at a particular position on the chromosome and may have different values, called “allele”.

In our problem, the chromosome, which codifies any trial feature vector (or individual, according to biological terminology), $F= [F_1, \ldots, F_p]$ is then a string $C= [c_1, \ldots, c_p]$, whose elements $c_i$ encode whether the $i$-th feature in the candidate vector $F$ is used or not. The allele $c_i = 1$ codifies that feature $F_i$ has been selected, while $c_i = 0$ indicates that this has not happened.

We complete this description by summarizing in Table 2 the main design parameters of the GA.

Table 2. Summary of the design parameters of the GA.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Initial population</td>
<td>100</td>
</tr>
<tr>
<td>Crossover probability</td>
<td>0.8</td>
</tr>
<tr>
<td>Mutation probability</td>
<td>0.1</td>
</tr>
<tr>
<td>Max. number of generations</td>
<td>100</td>
</tr>
<tr>
<td>Max number of iterations in which</td>
<td>20</td>
</tr>
<tr>
<td>MSE remains unchanged</td>
<td></td>
</tr>
</tbody>
</table>

4 Results

4.1 Experimental setup

To carry out the experiments, we have made use of the public database “The Berlin Database of Emotional Speech” [15]. This database is composed of 535 sound files, which have been generated by 10 difference actors (5 male and 5 female). As previously mentioned, there are seven different emotions included in the database: neutral, boredom, sadness, disgust, anxiety/fear, happiness, and anger. The language of the database is German.

As will be explained throughout this section, different experiments have been carried out as a function of the number of output emotions, the gender of the speakers or even more, by removing some speakers from the database. These batches of experiments carried out are explained in detailed in the paragraphs that follow.

- **Complete and edited database:** In the analysis of the preliminary results [14], we observed that the use of two particular subjects made the classification system behaved worse, or in other words, the error probability was higher. For this reason, we decided to edit the database by removing these two “particular” actors. In this sense, we have carried out a batch of experiments with the complete database (which includes all the actors), and with the edited database, in which the two worst actors have been removed.

- **Number of output emotions:**
  - Seven classes (seven emotions). The classification approach aims at automatically classifying among the seven emotions included in the database.
  - Three classes. In the analysis of the preliminary results [14], we obtained the confusion matrix shown in Table 3, in which we can observe that some emotions have a high percentage of confusion, like happiness and anger. Also anxiety and neutral are emotions may be highly confused. Bearing in mind also that the database may be unbalanced, we decided to group the seven emotions into three classes or levels of stress, as shown in “Group C1” and “Group C2” depicted in Table 4.

- **Gender classification:** We have carried out experiments for both genders, for only male speakers and for only female speakers.

![Table 3. Confusion matrix (%) Emotions: Neutral (N), Boredom (B), Sadness (S), Disgust (D), Anxiety/Fear (Ax), Happiness (H) and Anger (A).](image-url)
Finally, for properly completing this section, it is worth mentioning that since the database used in the experiments contains a low number of files, we have made use of the Bootstrap technique [13]. This strategy basically consists in three stages that are described below:

- To divide the database in two sets: a design set and a test set.
- To obtain the parameters of the classification system by using the design set.
- To obtain the performance of the classification system by using the test set.

These abovementioned steps are repeated 25 times by using a different design and test set for each iteration. Please note that in the design set are included four females and four males and in the test set are included the other female and male. For each iteration, the error probability is computed and after the 25 iterations, the mean error probability is obtained and that is just the result shown for each classification task.

4.2 Numerical results

Table 5 summarizes the experiments carried out in this work. As the reader can see, roughly speaking, two different batches of experiments have been carried out. The first one has been performed with the complete database, whereas the second one has been carried out with the edited database. Each experiment is subdivided into three different experiments, each sub-experiment corresponding to a different classification problem. These sub-experiments lead to a) seven emotions classification problem, and 2) three emotions classification problem ("Group C1") and 3) three emotions classification problem ("Group C2"). Each sub-experiment has been performed for both genders (male and female) and either only considering males or only considering females. With this in mind, a total of 18 classification tasks have been represented in Table 5. Please note that in all the mentioned classification tasks, it is used the GA-based feature selection approach described in Section 3 for the purpose of selecting the most appropriate features among the 57 features listed in Table 1. Once explained the experiments, it is seems clear to note that, at a first glance, better results are achieved by using the edited database. Deepening a little more in the results achieved by using the edited database, it is evident that better results are obtained when a lower number of output emotions are considered in the classification task. Regarding the results reached when using three emotions, the best result is achieved when "Group C1" is used for female speakers, leading to a probability error equal to 9.9%. Note that this classification performance is achieved by making use of 25 features.

Completing this section demands to mention that the average percentage of selected features for the 18 classification tasks has been found to be 36% for MFCCs, 21% for NHEV, 12% for HNR, 12% for DATA, 8% for SHIMMER, 5% for SPI, 5% for APQ and finally, 1% for VTI. Regarding these results, it is clear to note that the most selected features are MFCCs. It is also important to highlight that the percentage of selection of the feature proposed in this work, that is, NHEV, has been found to be 21%, which clearly exhibits that this novel feature works properly for the classification problem at hand.

Note the key importance of these results when compared to those achieved in subjective tests carried out in the laboratory. To be more precise, the mean percentage of error achieved by 10 subjects, discriminating among the seven emotions, was found to be approximately 40%.

<table>
<thead>
<tr>
<th>Emotions</th>
<th>Numer of Files Database</th>
<th>Stress Level</th>
<th>Group C1</th>
<th>Group C2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anger</td>
<td>127</td>
<td>High</td>
<td>High</td>
<td></td>
</tr>
<tr>
<td>Happiness</td>
<td>71</td>
<td>Medium</td>
<td>Medium</td>
<td></td>
</tr>
<tr>
<td>Disgust</td>
<td>69</td>
<td>Low</td>
<td>Low</td>
<td></td>
</tr>
<tr>
<td>Anxiety/fear</td>
<td>46</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sadness</td>
<td>81</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Boredom</td>
<td>62</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Neutral</td>
<td>79</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4. Distribution of emotions included in the database.

<table>
<thead>
<tr>
<th>Emotions</th>
<th>Gender</th>
<th>Low probability of error (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Seven</td>
<td>Both</td>
<td>31.2</td>
</tr>
<tr>
<td></td>
<td>Male</td>
<td>28.7</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>30.9</td>
</tr>
<tr>
<td>Three</td>
<td>Both</td>
<td>15.9</td>
</tr>
<tr>
<td>Group C1</td>
<td>Male</td>
<td>13.5</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>16.0</td>
</tr>
<tr>
<td>Three</td>
<td>Both</td>
<td>14.7</td>
</tr>
<tr>
<td>Group C2</td>
<td>Male</td>
<td>14.5</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>14.8</td>
</tr>
<tr>
<td>Seven</td>
<td>Both</td>
<td>27.2</td>
</tr>
<tr>
<td>Edited</td>
<td>Male</td>
<td>26.6</td>
</tr>
<tr>
<td>Database</td>
<td>Female</td>
<td>26.7</td>
</tr>
<tr>
<td>Three</td>
<td>Both</td>
<td>12.5</td>
</tr>
<tr>
<td>Group C1</td>
<td>Male</td>
<td>10.3</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>9.9</td>
</tr>
<tr>
<td>Three</td>
<td>Both</td>
<td>13.6</td>
</tr>
<tr>
<td>Group C2</td>
<td>Male</td>
<td>15.2</td>
</tr>
<tr>
<td></td>
<td>Female</td>
<td>11.7</td>
</tr>
</tbody>
</table>

Table 5. Classification approach that exhibits the lowest error probability.
5 Conclusion

The purpose of this paper is to obtain a system aiming at classifying emotions, and consequently, detecting stress levels, through voice analysis. In particular, we have aimed at discriminating among anger, happiness, disgust, anxiety/fear, sadness, boredom and neutral. In this regard, different experiments have been carried out in the effort of finding the best feature set. The results obtained have demonstrated that the feature set that exhibits best performance for the classification problem proposed in this work includes MFCCs, NHEV, DATA and HNR features. By discriminating among seven classes, the best classification performance has been found to be about 27% by using 30 features, whereas by grouping the seven emotions into three different classes, it has been observed that the error probability is decreased to 9.9% and making use of 25 features. It is important to note that the mentioned results are achieved by using the edited database, or in other words, the original database in which the two worst actors have been removed. Regarding the features selected, it is worth mentioning that the selection percentage of the feature proposed here, NHEV, has been found to be about 21%, only lower than that reached by MFCCs, which has been found to be 36%.

It is also important to highlight that the error probability obtained for male gender is lower than that reached by female gender when using the complete database. However, by using the edited database, the lowest error probability is achieved by using female gender. This basically means that it is interesting to study the results without some individuals, since it demonstrates that some people have much problems for expressing their feelings.

Note the key importance of these results when compared those achieved in subjective tests carried out in the laboratory. To be more precise, the mean percentage of error achieved by 10 subjects, discriminating among the seven mentioned emotions, was found to be about 40%.

Acknowledgments

This work has been funded by the Spanish Ministry of Education and Science (TEC2009-14414-C03-03), by the Spanish Ministry of Defense (DN8644-ATREC) and by the University of Alcala under project UAH2011/EXP-028.

References


