IEEE TRANSACTIONS ON
EMERGING TOPICS
IN COMPUTING

Received 10 April 2013; revised 17 July 2013; accepted 17 July 2013. Date of publication 25 July 2013;
date of current version 21 January 2014.

Digital Object Identifier 10.1109/TETC.2013.2274797

Gender-Driven Emotion Recognition Through Speech Signals for Ambient Intelligence Applications

IGOR BISIO, ALESSANDRO DELFINO, FABIO LAVAGETTO, MARIO MARCHESE, AND ANDREA SCIARRONE

Department of Electrical, Electronic, Telecommunications Engineering and Naval Architecture, University of Genoa, Genoa 16145, Italy
CORRESPONDING AUTHOR: I. BISIO (igor.bisio@unige.it)

ABSTRACT This paper proposes a system that allows recognizing a person's emotional state starting from audio signal registrations. The provided solution is aimed at improving the interaction among humans and computers, thus allowing effective human-computer intelligent interaction. The system is able to recognize six emotions (anger, boredom, disgust, fear, happiness, and sadness) and the neutral state. This set of emotional states is widely used for emotion recognition purposes. It also distinguishes a single emotion versus all the other possible ones, as proven in the proposed numerical results. The system is composed of two subsystems: 1) gender recognition (GR) and 2) emotion recognition (ER). The experimental analysis shows the performance in terms of accuracy of the proposed ER system. The results highlight that the *a priori* knowledge of the speaker's gender allows a performance increase. The obtained results show also that the features selection adoption assures a satisfying recognition rate and allows reducing the employed features. Future developments of the proposed solution may include the implementation of this system over mobile devices such as smartphones.

INDEX TERMS Human-computer intelligent interaction, gender recognition, emotion recognition, pitch estimation, support vector machine.

I. INTRODUCTION

Recently there has been a growing interest to improve human-computer interaction. It is well-known that, to achieve effective Human-Computer Intelligent Interaction (HCII), computers should be able to interact naturally with the users, i.e. the mentioned interaction should mimic human-human interactions. HCII is becoming really relevant in applications such as smart home, smart office and virtual reality, and it may acquire importance in all aspects of future peoples life. A peculiar and very important developing area concerns the remote monitoring of elderly or ill people. Indeed, due to the increasing aged population, HCII systems able to help live independently are regarded as useful tools. Despite the significant advances aimed at supporting elderly citizens, many issues have to be addressed in order to help aged ill people to live independently.

In this context recognizing people emotional state and giving a suitable feedback may play a crucial role. As a consequence, emotion recognition represents a hot research area in both industry and academic field. There is much research in this area and there have been some successful products [1]. Usually, emotion recognition systems are based on facial or voice features. This paper proposes a solution, designed to be employed in a Smart Environment, able to capture the emotional state of a person starting from a registration of the speech signals in the surrounding obtained by mobile devices such as smartphones.

Main problems to be faced concern: the concept of emotion, which is not precisely defined for the context of this paper; the lack of a widely accepted taxonomy of emotions and emotional states; the strong emotion manifestation dependency of the speaker. Emotion recognition is an extremely difficult task.

This paper presents the implementation of a voice-based emotion detection system suitable to be used over smartphone platforms and able to recognize six emotions (anger, boredom, disgust, fear, happiness, sadness) and the neutral state, as widely used for emotion recognition. Particular attention

is also reserved to the evaluation of the system capability to recognize a single emotion versus all the others. For these purposes, a deep analysis of the literature is provided and state-of-the-art approaches and emotion related features are evaluated. In more detail, to capture emotion information, 182 different features related to speech signals' prosody and spectrum shape are used; the classification task is performed by adopting the Support Vector Machine (SVM) approach.

The main contributions of this paper concern: *i)* a system able to recognize people emotions composed of two subsystems, Gender Recognition (GR) and Emotion Recognition (ER); *ii)* a gender recognition algorithm, based on pitch extraction, and aimed at providing *a priori* information about the gender of the speaker; *iii)* a SVM-based emotion classifier, which employs the gender information as input. Reduced feature sets, obtained by feature selection, performed through Principal Component Analysis (PCA), have been investigated and applied. In order to train and test the mentioned SVM-based emotion classifier, a widely used emotional database (called Berlin Emotional Speech Database, BESD) has been employed.

Experimental results show that the proposed system is able to recognize the emotional state of a speaker with an accuracy level often higher than the evaluated methods taken from the literature, without applying any pre-processing on the analysed speech signals. The obtained results show also that adopting a feature selection algorithm assures good recognition rate levels also when a consistent reduction of the used features is applied. This allow a strong limitation of the number of operations required to identify the emotional content of a particular audio signal. These peculiarities make the proposed solution suitable to operate on mobile platforms such as smartphones and tablets, in which the availability of computational resources and the energy consumption constitute issues of primary relevance.

The obtained results also show a strong dependency of the overall system reliability on the database adopted for training and testing phases: the use of a simulated database (i.e., a collection of emotion vocal expressions played by actors) allows obtaining a higher level of correctly identified emotions. In addition, the performed tests show that the SVM-based emotion classifier can be reliably used in applications where the identification of a single emotion (or emotion category) versus all the other possible ones is required, as in case of panic or annoyance detection.

II. RELATED WORK

If the phone is aware of its owner mood can offer more personal interaction and services. Mobile sensing, in recent years, has gone beyond the mere measure of physically observable events. Scientist studying affective computing [2], [3] have published techniques able to detect the emotional state of the user [2], [4]–[6] allowing the development of emotion-aware mobile applications [7]. Existing work focused on detecting emotions rely on the use of invasive means such as microphones and cameras [5], [6], [8], and body sensors worn by the user [7]. The proposed method based on the employment of audio signals represents an efficient alternative to the mentioned approaches. In the literature a traditional speech-based emotion recognition system consists of four principal parts:

- *Feature Extraction*: it involves the elaboration of the speech signal in order to obtain a certain number of variables, called features, useful for speech emotion recognition.
- *Feature Selection*: it selects the more appropriate features in order to reduce the computational load and the time required to recognize an emotion.
- *Database*: it is the memory of the classifier; it contains sentences divided according to the emotions to be recognized.
- *Classification*: it assigns a label representing the recognized emotion by using the features selected by the Feature Selection block and the sentences in the Database.

Given the significant variety of different techniques of Feature Extraction, Feature Selection, and Classification, and the breadth of existing databases, it is appropriate to analyse in detail each block.

A. FEATURES EXTRACTION

Many different speech feature extraction methods have been proposed over the years. Methods are distinguished by the the ability to use information about human auditory processing and perception, by the robustness to distortions, and by the length of the observation window. Due to the physiology of the human vocal tract, human speech is highly redundant and has several speaker-dependent features, such as pitch, speaking rate and accent. An important issue in the design of a speech emotion recognition system is the extraction of suitable features that efficiently characterize different emotions. Although there are many interesting works about automatic speech emotion detection [9], there is not a silver bullet feature for this aim.

Since speech signal is not stationary, it is very common to divide the signal in short segments called frames, within which speech signal can be considered as stationary. Human voice can be considered as a stationary process for intervals of 20–40 [ms]. If a feature is computed at each frame is called local, otherwise, if it is calculated on the entire speech is named global. There is not agreement in the scientific community on which between local and global features are more suitable for speech emotion recognition.

1) GENDER RECOGNITION FEATURES

Together with the Mel Frequency Cepstral Coefficients (MFCC) [10], pitch is the most frequently used feature [11]–[14] since it is a physiologically distinctive trait of a speaker's gender. Other employed features are formant frequencies and bandwidths, open quotient and source spectral tilt correlates [12], energy between adjacent formants [15], fractal dimension and fractal dimension

complexity [13], jitter and shimmer (pitch and amplitude micro-variations, respectively), harmonics-to-noise-ratio, distance between signal spectrum and formants [16].

2) EMOTION RECOGNITION FEATURES

Coherently with the wide literature in the field, in this paper a set of 182 features has been analysed for each the recorded speech signal, including:

- Mean, variance, median, minimum, maximum and range of the amplitude of the speech;
- Mean and variance of the speech Energy;
- Mean, variance, median, minimum, maximum and range of the pitch;
- Mean, variance, minimum, maximum and range of the first 4 formants;
- Energy of the first 22 Bark sub-bands [17];
- Mean, variance, minimum, maximum and range of the first 12 Mel-Frequency Cepstrum Coefficients [13], [14];
- Spectrum shape features: Center of Gravity, Standard Deviation, Skewness and Kurtosis;
- Mean and standard deviation of the glottal pulse period, jitter local absolute, relative average perturbation, difference of difference period and five-point period perturbation quotient.

B. FEATURES SELECTION AND REDUCTION

A crucial problem for all emotion recognition systems is the selection of the best set of features to characterize the speech signal. The purpose of this part is to appropriately select a subset of features from the original set in order to optimize the classification time and the accuracy. In the case of real-time applications reducing the number of used feature is crucial in order to limit the computational complexity and the required time to complete the emotion recognition process. An increase in classification performance usually would be expected when more features are used. Nevertheless, the performance can decrease for an increasing number of features if the number of patterns is too small. This phenomenon is known as the curse of dimensionality. This part also aims at reducing the speech features set size either by selecting the most relevant feature subset and removing the irrelevant ones or by generating few new features that contain most valuable speech information. The most performant strategy to get the best features set is an exhaustive search but it is often computationally impractical. Therefore, many sub-optimum algorithms have been proposed.

C. DATABASE

The database, also called dataset, is a very important part of a speech emotion recognizer. The role of databases is to assemble instances of episodic emotions. It is used both to train and to test the classifier and it is composed of a collection of sentences with different emotional content.

The most used are:

- *Reading-Leeds Database [18]*: project begun in 1994 to meet the need for a large, well-annotated set of natural or near-natural speeches orderly stored on computers. The essential aim of the project was to collect speeches that were genuinely emotional rather than acted or simulated.
- *Belfast Database*: it was developed as part of a project called Principled Hybrid Systems and Their Application (PHYSTA) [19], whose aim was to develop a system capable of recognizing emotion from facial and vocal signs.
- *CREST-ESP (Expressive Speech Database)*: database built within the ESP project [20]. Research goal was to collect a database of spontaneous, expressive speeches.
- *Berlin Emotional Speech (BES)*: this is the database employed in this paper. For this reason paragraph III-B.4 has been dedicated to it.

D. CLASSIFICATION METHODS

The last part is needed to train and build a classification model by using machine learning algorithms to predict the emotional states on the basis of the speech instances. The key task of this stage is to choose an efficient method to provide accurate predicted results for emotion recognition. Each classifier requires an initial phase in which it is trained to perform a correct classification and a subsequent phase in which the classifier is tested. There are several techniques to manage the two phases.

- Percentage split: the database is divided into two parts, used, respectively to train and to test the classifier.
- K-fold cross-validation [21]: it is a statistic technique usable when the training set contains many sentences. It allows mitigating the overfitting problem. In practice, the dataset is randomly divided into k parts of equal size. The algorithm acts in steps. At each step, one of these parts is used as test set while all the others are employed as training set. The procedure iterates until all the k parts have been used to test the classifier. Finally, the results of each step are averaged together.
- Leave-one-out cross-validation [22]: it is a variant of the K-fold cross-validation in which k is equal to the number of folds. In Leave-one-out cross-validation the class distributions in the test set are not related to the ones in the training data. Therefore it tends to give less reliable results. However it is still useful to deal with small datasets since it utilizes the greatest amount of training data from the dataset.
- The database is used both for the training phase and for the test phase.

Each classification method has advantages and drawbacks. Among the many available approaches, the most used are Maximum Likelihood Bayes (MLB) classifier [23], Support Vector Machine (SVM) [24], Hidden Markov Model (HMM) [9], Artificial Neural Network (ANN) [25], k-Nearest Neighbours (k-NN) [26]. Also other interesting classifiers

are used in a significant number of studies dedicated to the problem of speech emotion recognition and deserve to be referenced: Fuzzy Classifier [27], Decision Tree [28], Random Forest [29], Linear Discriminant Classifier (LDC) [30], Generative Vector Quantization (GVQ) [31].

E. PAPER CONTRIBUTIONS

The paper presents a gender-driven emotion recognition system whose aim, starting from speech recordings, is to individuate the gender of speakers and then, on the basis of this information, to classify the emotion characterizing the speech signals.

Concerning the first step, the paper proposes a gender recognition method based on the pitch. This method employs a typical speech signal feature and a novel extraction method. It guarantees excellent performance: 100% accuracy. In practice, it always recognises the gender of the speaker.

Concerning the emotion recognition approach, the paper proposes a solution based on traditional features sets and classifiers but, differently from the state of the art, it employs two classifiers (i.e., two Support Vector Machines): the one trained on the basis of signals recorded by male speakers and the other one trained by female speech signals. The choice between the two classifiers is driven by the gender information individuated through the gender recognition method.

To the best of authors' knowledge, the proposed gender-driven emotion recognition system represents a novel approach with respect to the literature in the field.

III. GENDER-DRIVEN EMOTION RECOGNITION SYSTEM ARCHITECTURE

The system is aimed at recognizing 7 different emotions: anger, boredom, disgust, fear, happiness, sadness, and neutral state. The overall system scheme is reported in Fig. 1.

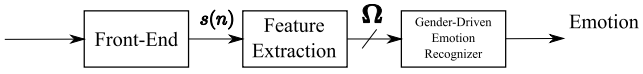

FIGURE 1. Proposed emotion recognition scheme overall architecture.

The quantity $s(t)$ represents the original continuous input audio signal. The Front-End block acquires $s(t)$ and samples it with frequency $F_S = 16$ $[KHz]$ in order to obtain the discrete sequence $s(n)$. After this step, a feature vector $\mathbf{\Omega}$ is computed by the Features Extraction block. It is worth noticing that $\mathbf{\Omega}$ includes the features $\mathbf{\Omega}^{GR}$ and $\mathbf{\Omega}^{ER}$ respectively employed by the Gender Recognition and the Emotion Recognition subsystems. In practice the feature vector may be written as $\mathbf{\Omega} = \{\mathbf{\Omega}^{GR}, \mathbf{\Omega}^{ER}\}$. $\mathbf{\Omega}$ is employed by the Gender-driven Emotion Recognition block that provides the output of the overall process: the recognized emotion. As already said and discussed in the reminder of this Section, this block is divided into two subsystems: Gender Recognition (GR) and Emotion Recognition (ER).

A. GENDER RECOGNITION (GR) SUBSYSTEM

As reported in papers such as [10], [11], [13], [32], [33] audio-based Gender Recognition (GR) has many applications. For example: gender-dependent model selection for the improvement of automatic speech recognition and speaker identification, content-based multimedia indexing systems, interactive voice response systems, voice synthesis and smart human-computer interaction. In this paper, the recognition of the gender is used as input for the emotion recognition block. As shown in the numerical result section, this pre-filtering operation improves the accuracy of the emotion recognition process.

Different kinds of classifiers are used to identify the speaker gender starting from features: e.g., Continuous Density Hidden Markov Models [13], [16], Gaussian Mixture Model (GMM) [14], [16], Neural Networks [10], [32], Support Vector Machines [12], [33]. The percentages of correct recognition of the speaker gender are reported in Table 1 for most classifiers referenced above.

TABLE 1. Classification accuracy (percentage) obtained by the evaluated GR methods.

Reference	Accuracy
[10]	100
[33]	100
[14]	98.35
[12]	95
[32]	91.7
[13]	90.9
[34]	90

1) PROPOSED GENDER RECOGNITION ALGORITHM

The proposed GR method is designed to distinguish a male from a female speaker and has been thought to be realized over mobile devices, such as smartphones. It is designed to operate in an open-set scenario and is based on audio pitch estimation. In a nutshell: it is based on the fact that pitch values of male speakers are on average lower than pitch values of female speakers because male vocal folds are longer and thicker compared to female ones. In addition, being male and female pitch frequency separated, we realized that satisfying results in terms of accuracy of the GR can be obtained by using a single-feature threshold γ_{thr} classifier rather than more complex and time-consuming ones. Furthermore, being mobile devices the target technology, time constraints must be carefully considered in the design, in particular in view of real-time applications.

The chosen feature is the mean of the Probability Density Function (PDF), whose definition is reported in Section III-A.3, of a number of frames of the voice signal, as explained below.

The signal to be classified as "Male" or "Female" is identified as $s(n), n = 1 \ldots N$. The GR method introduced in this paper is composed of the following steps:

1) The signal $s(n)$ is divided into frames.
2) The pitch frequency for each frame is estimated.
3) A number of frames of $s(n)$ is grouped into a odd-number of blocks.
4) The pitch PDF is estimated for each block.
5) The mean of each pitch PDF (PDF_{mean}) is computed.
6) The decision about "Male" or "Female" is taken, for each block, by comparing their PDF_{mean} with a fixed threshold γ_{thr} computed by using the training set.
7) The final decision on the whole signal gender is taken by the majority rule: the signal $s(n)$ is classified as "Male" if the majority of its blocks are classified as "Male". Otherwise, it is classified as "Female".

Average pitch frequencies for male and female speakers and the individuated threshold γ_{thr}, referred to a recording of 20 blocks are shown in Fig. 2.

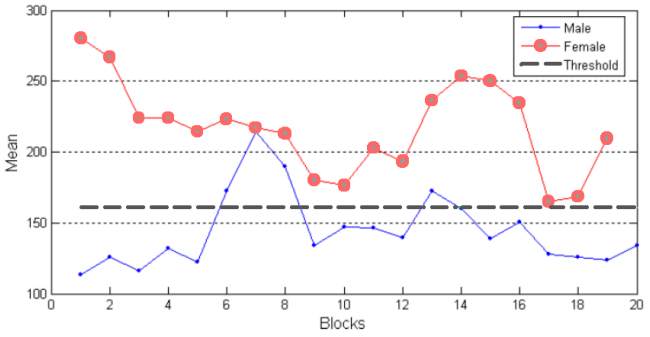

FIGURE 2. Average pitch frequencies, referred to male and female single speakers, and the employed threshold γ_{thr}, for a recording divided in 20 blocks.

The proposed GR method has a very low computational complexity and therefore consumes a limited quantity of energy, nevertheless it guarantees 100% recognition performance, as the solutions proposed in [10] and [33].

The reminder of this section is focused on the detailed description of the single steps followed by the GR algorithm.

2) PITCH FREQUENCY ESTIMATION

Speech signal exhibits a relative periodicity and its fundamental frequency, called pitch (frequency), is usually the lowest frequency component [35]. In the case of this paper, it is important to estimate the Probability Density Function (PDF) of the pitch of a given speech signal. The applied procedure applied will be described in the following subsection. In general, for voice speech, pitch is usually defined as the rate of vibration of the vocal folds [36] and for this reason, can be considered a distinctive feature of an individual. Estimating the pitch of an audio sample could therefore help classify it as belonging to either a male or a female, since its value for male speakers is usually lower than the one for female speakers. Many pitch estimation algorithms have been proposed in the past years, involving both time- and frequency-domain analysis [35]. Many developed methods are context-specific, but pitch estimators designed for a particular application depend on the data domain and are typically less accurate when applied to a different domain. A method based on the signal autocorrelation has been chosen because of its good applicability to voice and ease of implementation. This method has been further adjusted by considering the physiological nature of voice signals and by downsampling the autocorrelation function, as described below.

In particular, given a real-value discrete-time signal $s(n)$, $n \in [1 \dots N]$ we have:

$$R(\tau) = \sum_{n=0}^{N-1} s(n)s(n+\tau) \qquad \tau \in [0, 1 \dots N-1] \tag{1}$$

$R(\tau)$ in (1) is the autocorrelation of lag τ. For the specific case of this paper, which deals with audio speech signals, the set of possible samples τ of the autocorrelation function can be reduced. [37] reports that the pitch of a speech signal, due to physiological reasons, is contained in a limited range $[P_1, P_2]$ (typically $P_1 = 50$ $[Hz]$ and $P_2 = 500$ $[Hz]$) and limits the τ range between τ_1 and τ_2, defined in (2).

$$\tau_1 = \left\lfloor \frac{F_s}{P_2} \right\rfloor \text{ and } \tau_2 = \left\lfloor \frac{F_s}{P_1} \right\rfloor \tag{2}$$

F_s is the sampling frequency applied to the original analog signal to obtain the discrete-time signal $s(n)$. In practice, the applied autocorrelation is defined in (3):

$$\hat{R}(\tau) = \sum_{n=0}^{N-1-\tau} s(n)s(n+\tau) \qquad \tau \in [\tau_1, \tau_1+1, \tau_1+2 \dots \tau_2]. \tag{3}$$

From the practical viewpoint, the physiological limitation of the pitch range implies a first significant reduction of the number of samples involved in the computation and, as a consequence, of the overall complexity. The discrete-time signals $s(n)$ is divided into frames, in this paper each of them composed of $N = 640$ samples, acquired by using $F_s = 16$ $[KHz]$. The value of N is chosen in order to obtain frames of $L_f = 40$ $[ms]$, which allow considering human voice as a stationary statistical process.

If the entire autocorrelation function $R(\tau)$ computed as in (1) is evaluated, the number of samples τ is equal to 640. Considering the physiological pitch limitations reported above, (i.e., $P_1 = 50$ $[Hz]$ and $P_2 = 500$ $[Hz]$), $\tau_1 = 32$ and $\tau_2 = 320$, as indicated in (2), $\hat{R}(\tau)$ in (3) is calculated by using $\tau_2 - \tau_1 + 1 = 289$ samples.

The autocorrelation function shows how much the signal correlates with itself, at different delays τ. Considering that, given a "sufficiently periodical" speech recording, its autocorrelation will present the highest value at delays corresponding to multiples of pitch periods [35], [37].

Defining the pitch period as

$$\tau_{pitch} = \arg\max_{\tau} \hat{R}(\tau) \tag{4}$$

the frequency of pitch is computed as

$$\rho_{pitch} = \frac{F_s}{\tau_{pitch}}. \tag{5}$$

To further reduce the computational complexity of the proposed pitch estimation method, a downsampled version of the autocorrelation function is introduced in this paper by using a downsampling factor $r < 1$ and $\frac{1}{r} \in \mathbb{N}$. Being N the cardinality of the original set of autocorrelation samples, the downsampled version uses $K = rN$ samples. In practice, the downsampled autocorrelation is defined as:

$$\begin{aligned} \tilde{R}(\tau) &= \sum_{n=0}^{N-1-\tau} s(n)s(n+\tau) \\ \tau &\in \left[\tau_1, \tau_1 + \tfrac{1}{r}, \tau_1 + \tfrac{2}{r} \ldots \tau_2\right]. \end{aligned} \tag{6}$$

It means that $\tilde{R}(\tau)$ considers just one sample of $\hat{R}(\tau)$ out of $\frac{1}{r}$ in the interval $[\tau_1 \ldots \tau_2]$.
As a consequence (4) and (5) become:

$$\tilde{\tau}_{pitch} = \arg\max_{\tau} \tilde{R}(\tau) \tag{7}$$

$$\tilde{\rho}_{pitch} = \frac{F_s}{\tilde{\tau}_{pitch}}. \tag{8}$$

In order to still correctly determine the maximum of the autocorrelation in (3), thus preventing errors in pitch estimation, a maximum *"Fine Search"* method has been designed and implemented in this paper to partially compensate the inaccuracies introduced by downsampling. Starting from the delay corresponding to the pitch $\tilde{\tau}_{pitch}$, obtained by the downsampled autocorrelation function, the values of $\hat{R}(\tau)$, in (3), are computed for all the τ values adjacent to $\tilde{\tau}_{pitch}$ up to a depth of $\pm|\frac{1}{r} - 1|$. Their maximum is taken as new pitch period τ'_{pitch}. Analytically:

$$\begin{aligned} \tau'_{pitch} &= \arg\max_{\tau} \hat{R}(\tau) \\ &\tau \in \left[\tilde{\tau}_{pitch} - \frac{1}{r} + 1 \ldots \tilde{\tau}_{pitch} - 1, \tilde{\tau}_{pitch}, \right. \\ &\left. \tilde{\tau}_{pitch} + 1 \ldots \tilde{\tau}_{pitch} + \frac{1}{r} - 1\right] \end{aligned} \tag{9}$$

$$\rho'_{pitch} = \frac{F_s}{\tau'_{pitch}} \tag{10}$$

ρ'_{pitch} is the reference pitch value in the remainder of this paper. Autocorrelation downsampling may cause the ''Fine Search'' to be applied around a local maximum of the autocorrelation in (3) instead of the global one. This occurs only if the delay corresponding to the global maximum of $\tilde{R}(\tau)$ is farther than $\pm|\frac{1}{r} - 1|$ samples from the delay corresponding to the global maximum of $\tilde{R}(\tau)$. This event occurs rarely and, as a consequence, the new approach guarantees a good estimation and represents a reasonable compromise between performance and computational complexity energy saving, making feasible the implementation of the GR algorithm over mobile platforms.

3) PDF ESTIMATION

Looking at the list of steps in III-A.1 but providing more detail, the signal $s(n), n = 1 \ldots N$ is divided into $F = \lfloor \frac{N}{L} \rfloor$ frames, where L is the number of the sample in each frame. The value of L directly derives from $L = F_s \cdot L_f$. Generic i-th frame is defined as

$$f_i = \{s(n) : n = (i-1)L + 1, ..., iL\}, \quad i = 1, ..., F. \tag{11}$$

A pitch estimate is computed for each frame by applying the method described in Section III-A.2. Sets of consecutive frames are grouped together in blocks, in order to allow the computation of a pitch PDF for each block. Consecutive blocks are overlapped by V frames (i.e., the last V frames of a block are the first V frames of the following one) in order to take into account the possibility that a signal portion representing a speech falls across consecutive blocks if blocks were not overlapped. The signal contribution to the classification process would be divided between two separate blocks. The overlap implies there are $B = \lfloor \frac{F-V}{D-V} \rfloor$ blocks. The t-th block can be defined as

$$\begin{aligned} b_t = \{&f_i : i = (t-1)(D-V) + 1, ..., \\ &tD - (t-1)V\}, \quad t = 1, ..., B. \end{aligned} \tag{12}$$

For each b_t block there are V pitch values computed as in (10) identified as $\rho_{pitch}^{b_t,v}, v = 1, ..., V$. b_t block PDFs span over a frequency interval ranging from the minimum to the maximum computed pitch value. Such frequency interval is divided into H smaller frequency bins of Δp [Hz] size determined through extensive tests. Being p is the variable identifies the frequency the PDF for each block b_t is estimated by a weighted sum of the number of occurrences of single $\rho_{pitch}^{b_t,v}, v = 1, ..., V$ within each frequency bin $h = 0, ..., H-1$. In formula:

$$PDF(p) = \sum_{h=0}^{H-1} w_h \cdot rect\left(\frac{p - \left[\left(\frac{1}{2} + h\right)\Delta p\right]}{\Delta p}\right) \tag{13}$$

w_h is the coefficient associated to the h-th bin and implements the mentioned weighted sum, as explained in the following. If w_h is the number of $\rho_{pitch}^{b_t,v}, v = 1, ..., V$, whose values fall within the h-th bin, then the PDF is simply estimated through the number of occurrences and is called ''histogram count''. In order to have a more precise PDF estimation and, consequently, more accurate features vectors, this paper links the coefficient w_h to the energy distribution of the signal $s(n)$ in the frequency range of the PDF.

Given the Discrete Fourier Transform (DFT) of $s(n)$, $DFT(s(n)) = S(k) = \sum_{n=0}^{N-1} s(n) \cdot e^{-j2\pi n\frac{k}{N}}, \forall k \in [0, ..., N-1]$, the signal energy ($E_s$) definition and the *Parseval Relation*, written for the defined DFT, we have $E_s = \sum_{n=0}^{N-1} |s(n)|^2 = \frac{1}{N}\sum_{k=0}^{N-1} |S(k)|^2$. A single energy component is $|S(k)|^2$, where k represents the index of the frequency $f_k = \frac{k}{N}$, $k = 0, ..., N-1$. To evaluate the energy component of each frequency bin h, we would need to know the energy contribution carried by each pitch occurring within bin h. In practice, the quantity $|S(\rho_{pitch}^{b_t,v})|^2, v = 1, ..., V$ would be necessary but the DFT is a function of an integer number k and $\rho_{pitch}^{b_t,v} \in \mathbb{R}$. So, to allow the computation for real numbers,

$\rho_{pitch}^{b_t,v}$ is approximated by the closest integer number $\rho_{pitch,int}^{b_t,v}$, defined as follows:

$$\rho_{pitch,int}^{b_t,v} = \begin{cases} \left\lfloor \rho_{pitch}^{b_t,v} \right\rfloor & \text{if } \rho_{pitch}^{b_t,v} - \left\lfloor \rho_{pitch}^{b_t,v} \right\rfloor < \frac{1}{2} \\ \left\lceil \rho_{pitch}^{b_t,v} \right\rceil & \text{if } \rho_{pitch}^{b_t,v} - \left\lfloor \rho_{pitch}^{b_t,v} \right\rfloor \geq \frac{1}{2} \end{cases} \tag{14}$$

Consequently, the coefficient w_h, properly normalized, is defined as

$$w_h = \frac{\sum_{v=1}^{V} \left| S(\rho_{pitch,int}^{b_t,v}) \right|^2}{\sum_{h=0}^{H-1} \sum_{v=1}^{V} \left| S(\rho_{pitch,int}^{b_t,v}) \right|^2} \tag{15}$$

The sums with index v must be computed for each $\rho_{pitch}^{b_t,v} \in bin\ h$. Directly from (15), higher-energy pitch estimates become more relevant than lower-energy ones. Actually the underlying hypothesis is that higher-energy pitches derive from fully voiced (without silence intervals) frames and are therefore more reliable. This section leads to more distinct PDFs and more accurate features vector, thus significantly improving the GR method performance compared to computing PDFs by simply executing a ''histogram count'', for which w_h is the number of pitches whose value falls within the h-th bin.

4) FEATURES VECTOR DEFINITION AND GENDER CLASSIFICATION POLICY

In order to determine the best feature vector $\mathbf{\Omega}^{GR}$ that maximizes the efficiency of the proposed GR method, different feature vectors were evaluated by combining different individual features:

- PDF maximum: PDF_{max};
- PDF mean: PDF_{mean};
- PDF standard deviation: PDF_{std};
- PDF roll-off: $PDF_{roll-off}$;

By using a general set of features, the feature vector would be composed, for each block, by

$$\mathbf{\Omega}^{GR} = \left\{ \omega_1^{GR}, \cdots, \omega_z^{GR}, \cdots, \omega_Z^{GR} \right\} \tag{16}$$

where $z \in [1, Z]$ and Z is the size of the defined features vector. In this paper, each element of $\mathbf{\Omega}^{GR}$ is one of the features extracted from the Pitch PDF listed above so that $\mathbf{\Omega}^{GR} = \left\{ \omega_1^{GR}, \omega_2^{GR}, \omega_3^{GR}, \omega_4^{GR} \right\}$.

For sake of completeness, different subsets of the features listed above have been tested as features vector $\mathbf{\Omega}^{GR}$. However, due to the separation between male and female pitch frequency, practical experiments have shown that the employment of the *PDFmean* is sufficient to separate the two classes. For this reason the feature vector is reduced to a simple scalar $\mathbf{\Omega}^{GR} = \omega_1^{GR} = PDFmean$.

From an analytical viewpoint, *PDFmean* is computed as follows:

$$\mathbf{\Omega}^{GR} = \omega_1^{GR} = PDF_{mean} = \sum_{h=0}^{H-1} p_h \cdot w_h \tag{17}$$

where p_h represents the central frequency of the h-th bin and w_h is computed as in (15). The label g of the recognized gender is obtained through (18). g has value 1 for the Male and -1 for Female.

$$\begin{aligned} g = g(\mathbf{\Omega}^{GR}) = g(\omega_1^{GR}) = -sgn(\omega_1^{GR} - \gamma_{thr}) = \\ = -sgn\left(\sum_{h=0}^{H-1} p_h \cdot w_h - \gamma_{thr} \right) \end{aligned} \tag{18}$$

In practice, $g = 1$ if $\sum_{h=0}^{H-1} p_h \cdot w_h \leq \gamma_{thr}$, $g = -1$ otherwise. Starting from experimental tests, the employed threshold γ_{thr} has been estimated to be 160 [Hz]. This numerical value has been employed in the proposed gender-driven emotion recognition system.

Numerical results, not reported here for sake of brevity, have shown that the proposed method is able to recognize the gender with 100% accuracy.

B. EMOTION RECOGNITION (ER) SUBSYSTEM

The implemented Emotion Recognition (ER) subsystem is based on two inputs: the features extracted by the Features Extraction Block $\mathbf{\Omega}$, in particular the sub-set $\mathbf{\Omega}^{ER}$ of features needed for the emotion recognition and the recognized speaker gender provided by the GR subsystem. Differently from the the GR subsystem in which the employed feature has been individuated (the Pitch), concerning the ER subsystem the selection of feature(s) to be employed is still an open issue. For this reason, this paper does not provide a fixed set of features but proposes a study that takes into account the most important features employed in the literature and their selection through a features selection algorithm. Indeed, the features employed in the ER subsystem are based on a set of features (182 in this paper) or on a sub-set of them. Subsets have been individuated by using a *Principal Component Analysis* (PCA) algorithm and have been evaluated in terms of recognition rate. The recognition rate obtained by varying the selected features has been reported in Section IV.

1) PRINCIPAL EMOTION FEATURES

For the sake of completeness, in the reminder of this subsection, the definitions of the considered principal features are listed and defined. Energy and amplitude are simply the energy and the amplitude of the audio signals and no explicit formal definition is necessary. Concerning *Pitch* related features the extraction approach is based on the pitch estimation described in Section III-A.2.

The other considered features can be defined as follows:

a) Formants: In speech processing, formants are the resonance frequencies of the vocal tract. The estimation of their frequency and their -3 [dB] bandwidth is fundamental for the

analysis of the human speech as they are meaningful features able to distinguish the vowel sounds.

In this paper, we employ a typical method to compute them, which is based on the Linear Predictive Coding (LPC) analysis. In more detail, the speech signal $s(n)$ is re-sampled at twice the value $F_{max} = 5.5$ $[Hz]$, which is the maximum frequency applied within the algorithm to search formants. Then, a pre-emphasis filter is applied. The signal is divided in audio frames (0.05 $[s]$ long in the case of formant extraction) and a Gaussian window is applied to each frame. After that LPC Coefficients are computed by using the Burg method [38]. Being $z_i = r_i e^{\pm\theta_i}$ the i-th complex root pair of the prediction (LPC) polynomial, the frequency, called Υ_i, and the -3 [dB] bandwidth, indicated with Δ_i, of the i-th formant related to the i-th complex root pair of the LPC polynomial, can be estimated by applying the following formulae [39]:

$$\Upsilon_i = \frac{F_s}{2\pi}\theta_i \tag{19}$$

$$\Delta_i = -\frac{F_s}{\pi}\ln r_i \tag{20}$$

The algorithm finds all the formants in the range $[0 - F_{max}]$ $[Hz]$. Some artefacts of the LPC algorithm can produce ''false'' formants near 0 and F_{max} $[Hz]$ therefore the formants below 50 and over $(F_{max} - 50[Hz])$ are removed.

b) Mel-Frequency Cepstrum Coefficients: Mel-Frequency Cepstrum (MFC) is a widely used representation of the short-term power spectrum of a sound (i.e., an audio signal). Mel-Frequency Cepstral Coefficients (MFCCs) allow describing MFC. MFCCs computation is based on the subdivision of the audio signal into analysis frames, whose duration is 30 $[ms]$, selected so that the distance between the centres of two adjacent frames is equal to 10 $[ms]$ (i.e., two consecutive frames are overlapped for one third of their duration). For each frame the spectrum is shaped through the so called ''mel-scale'', using 13 triangular overlapping filters. Finally, the MFCCs of each frame are computed by using the Discrete Cosine Transform (DCT).

$$MFCC_i = \sum_{j=1}^{13} P_j \cos\left(\frac{i\pi}{13}(j-0.5)\right), \quad i \in [1, M] \tag{21}$$

where P_j represent the power, in $[dB]$, of the output the j-th filter and M is the number of considered MFFCs. In our case, $M = 12$.

c) Center of Gravity: The spectral Centre Of Gravity (COG) is a measure of how high the frequencies in a spectrum are. For this reason the COG gives an average indication of the spectral distribution of the speech signal under observation. Given the considered discrete signal $s(n)$ and its DFT $S(k)$, the COG has been computed by:

$$f_{COG} = \frac{\sum_{k=1}^{N} f_k \, |S(k)|^2}{\sum_{k=1}^{N} |S(k)|^2} \tag{22}$$

where, as defined in Section III-A.3, $f_k = \frac{k}{N}$, $k = 0, ..., N-1$ represents the k-th frequency composing the DFT.

d) Spectrum Central Moments: The m-th central spectral moment of the considered sequence $s(n)$ has been computed by:

$$\mu_m = \frac{\sum_{k=1}^{N} (f_k - f_{COG})^m \, |S(k)|^2}{\sum_{k=1}^{N} |S(k)|^2} \tag{23}$$

e) Standard Deviation (SD): The standard deviation of a spectrum is defined as the measure of how much the frequencies in a spectrum can deviate from the centre of gravity. SD corresponds to the square root of the second central moment μ_2:

$$\sigma = \sqrt{\mu_2} \tag{24}$$

f) Skewness: The skewness of a spectrum is a measure of symmetry and it is defined as the third central moment of the considered sequence $s(n)$, divided by the 1.5 power of the second central moment:

$$\gamma_1 = \frac{\mu_3}{\sqrt{\mu_2^3}} \tag{25}$$

g) Kurtosis: The excess of Kurtosis is defined as the ratio between the fourth central moment and the square of the second central moment of the considered sequence $s(n)$ minus 3:

$$\gamma_2 = \frac{\mu_4}{\mu_2^2} - 3 \tag{26}$$

h) Glottal Pulses: Human speech, in the time domain, presents a periodic pattern. Each of the identifiable repeating patterns is called ''cycle'' and each peak in a cycle is called ''glottal pulse''. The duration of each cycle is called ''period of the glottal pulse''. The Dynamic Waveform Matching (DWM) has been used to find Glottal Pulses. The mean of the glottal pulse period ($\bar{T}$) is defined as follows:

$$\bar{T} = \frac{1}{Q}\sum_{q=1}^{Q} T_q \tag{27}$$

where T_q is the duration of the q-th glottal pulse period and Q is the number of periods. Among the Glottal Pulses, only the following have been used:

- Jitter Local Absolute (J_{LA}) is often used as a measure of voice quality. J_{LA} is defined as the average absolute

difference between consecutive glottal pulses intervals:

$$J_{LA} = \frac{1}{Q-1}\sum_{q=2}^{Q}\left|T_q - T_{q-1}\right| \tag{28}$$

- Relative average perturbation (J_{RAP}) is a jitter measure between an interval and its two neighbours:

$$J_{RAP} = \frac{1}{(Q-1)\bar{T}}\sum_{q=2}^{Q-1}\left|T_q - \frac{T_{q-1}+T_q+T_{q+1}}{3}\right| \tag{29}$$

- Difference of Difference Period (J_{DDP}) is defined as follows:

$$J_{DDP} = \frac{1}{(Q-2)\bar{T}}\sum_{q=2}^{Q-1}(T_{q+1}-T_q)(T_q - T_{q-1}) \tag{30}$$

- Five-points period perturbation quotient (J_{5PPQ}) is a jitter measure between an interval T_q and the average of T_q and its four closest neighbours: T_{q-2}, T_{q-1}, T_{q+1} and T_{q+2}.

$$J_{5PPQ} = \frac{1}{(Q-4)\bar{T}}\sum_{q=3}^{Q-3}\left|T_q - \frac{T_{q-2}+T_{q-1}+T_q+T_{q+1}+T_{q+2}}{5}\right| \tag{31}$$

2) EMOTION FEATURES SELECTION

As a preliminary step, the Principal Components Analysis (PCA) algorithm has been used in order to reduce and limit the number of features. PCA is a technique that, given high-dimensional feature vectors reduces the number of features used in the vector without losing too much information, by using the dependencies between features and by identifying the principal directions in which the features vary.

From each feature involved in the algorithm the mean value is subtracted, in order to obtain zero-mean feature set. After this step, PCA computes new variables called Principal Components (PCs) which are obtained as linear combinations of the original features. The first PC is required to have the largest possible variance (which, considering zero-mean feature sets, is equivalent to the concept of inertia, originally employed in the PC definition, see [40]). The second component is computed under the constraint of being orthogonal to the first one and to have, again, the largest possible variance. The other components are computed in a similar way. The values of PCs for the features are called *factor scores* and can be interpreted geometrically as the projections of the features onto the principal components [40]. The importance of a feature for a PC is directly proportional to the correspondent squared *factor score*. This value is called the *contribution* of the feature to the PC.

The value of the *contribution* is between 0 and 1 and, for a given PC, the sum of the *contributions* of all features is equal to 1. From a practical viewpoint, larger the value of the *contribution*, bigger the feature contributes to the PC.

Contribution values will be used in the numerical results section, in order to evaluate the Emotion Recognition block when different number of features are employed, starting from the one which presents the highest *contribution*.

3) EMOTIONS CLASSIFIERS

Usually, in the literature of the field, a Support Vector Machine (SVM) is used to classify sentences. SVM is a relatively new machine learning algorithm introduced by Vapnik [24] and derived from statistical learning theory in the 90s. The main idea is to transform the original input set into a high-dimensional feature space by using a kernel function and, then, to achieve optimum classification in this new feature space, where a clear separation among features obtained by the optimal placement of a separation hyperplane under the precondition of linear separability.

Differently from the previously proposed approaches, two different classifiers, both kernel-based Support Vector Machines (SVMs), have been employed in this paper as shown in Figure 3.

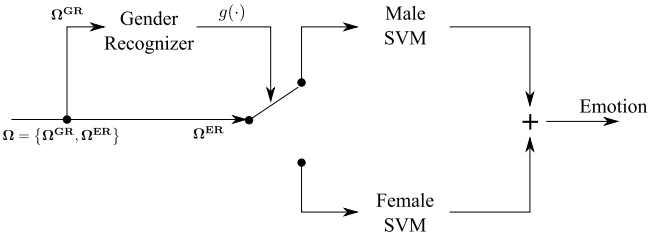

FIGURE 3. Emotion Recognition (ER) subsystem.

The first one (called Male-SVM) is used if a male speaker is recognized by the Gender Recognition block. The other SVM (Female-SVM) is employed in case of female speaker. Male-SVM and Female-SVM classifiers have been trained by using speech signals of the employed reference DataBase (DB) generated, respectively, by male and female speakers.

Being $g = \{1, -1\}$ the label of the gender as defined in Section III-A.4, the two SVMs have been trained by the traditional Quadratic Programming (QP) as done in [41]. In more detail, the following problem has been solved for each gender g:

$$\begin{aligned}\min_{\boldsymbol{\lambda}_g}\Gamma_g(\boldsymbol{\lambda}_g) &= \frac{1}{2}\sum_{u=1}^{\ell_g}\sum_{v=1}^{\ell_g} y_g^u y_g^v \phi(\boldsymbol{x}_g^u, \boldsymbol{x}_g^v)\lambda_g^u\lambda_g^v - \sum_{u=1}^{\ell_g}\lambda_g^u,\\ &\sum_{u=1}^{\ell_g}\lambda_g^u y_g^u = 0,\\ &0 \le \lambda_g^u \le C,\ \forall u\end{aligned} \tag{32}$$

where $\boldsymbol{\lambda}_g = \{\lambda_g^1 \ldots \lambda_g^u \ldots \lambda_g^{\ell_g}\}$ represents the well-known *Lagrangian Multipliers* vector of the QP problem written in dual form. Vectors $\boldsymbol{x}_g^1, \ldots, \boldsymbol{x}_g^u, \ldots, \boldsymbol{x}_g^{\ell_g}$ are features vectors while scalars $y_g^1, \ldots, y_g^u, \ldots, y_g^{\ell_g}$ are related labels (i.e., the emotions in this paper). They represent the vectors of the training set for the g-th gender. $(\boldsymbol{x}_g^u, y_g^u)$, $\forall u \in [1, \ell_g]$ is the related association, also called *observation*, between the u-th input features vector $\boldsymbol{x}_g^u$ and its label y_g^u. The quantity ℓ_g

is the total amount of *observations* composing the training set. The quantity C ($C > 0$) is the Complexity constant which determines the trade-off between the flatness (i.e., the sensitivity of the prediction to perturbations in the features) and the tolerant level for misclassified samples. Higher value of C means that is more important minimising the degree of misclassification. $C = 1$ is used in this paper.

Equation (32) represents a non-linear SVM and the function $\phi(\boldsymbol{x}_g^u, \boldsymbol{x}_g^v)$ is the Kernel function that, in this paper, is $\phi(\boldsymbol{x}_g^u, \boldsymbol{x}_g^v) = (\boldsymbol{x}_g^u)^T(\boldsymbol{x}_g^v) + 1$.

Coherently with [41], the QP problems (one for each gender) in equation (32) are solved by the *Sequential Minimal Optimization* (SMO) approach that provides an optimal point, not necessarily unique and isolated, of (32) if and only if *Karush-Kuhn-Tucker* (KKT) conditions are verified and matrices $y_g^u y_g^v \phi(\boldsymbol{x}_g^u, \boldsymbol{x}_g^v)$ are positive semi-definite. Details about the KKT conditions and the SMO approach employed to solve problem (32) can be found in [41] and references therein.

4) EMPLOYED SIGNAL DATASET: BERLIN EMOTIONAL SPEECH (BES)

As described in [42], and here reported for the sake of completeness, BES is a public database of acted speeches. The sentences are recorded by 10 German actors (5 male and 5 female) that produced 10 utterances each (5 short and 5 long phrases). Each utterance is classified with one among 7 different labels: anger, boredom, disgust, fear, happiness, sadness, and neutral. The sentences were evaluated by 20 listeners to check the emotional state and only those that had a recognition rate of 80% or above were retained, getting about 500 speeches. Additionally, two more perception tests were carried out: one to rate the strength of the displayed emotion for each speech, the other to judge the syllable stress of every speech. Emotional strength was used as a control variable in statistical analysis. Evaluating the syllable stress was necessary because objective stress measurements are not available. This last test was performed only by phonetically trained subjects. Speeches were recorded within an anechoic chamber with high-quality recording equipment. Recordings are sampled at 16 [KHz].

IV. PERFORMANCE EVALUATION

In this Section the performance evaluation of the overall Emotion Recognition (ER), in terms of accuracy (i.e., correct detection rate), of the system is presented. The recognized emotions are: anger (AN), boredom (BO), disgust (DI), fear (FE), happiness (HA), sadness (SA), together with the neutral (NE) state. The reported results are divided into two main parts. The first part shows the performance of the system if no information about the gender of the speaker is exploited in the emotion recognition process. The second part of the results provides the performance obtained by exploiting the knowledge related to the speakers' gender. The experimental results highlight that the gender information allows incrementing the accuracy of the emotion recognition system on average.

A. WITHOUT GENDER RECOGNITION

In this subsection, the accuracy of a traditional approach, without having any "*a priori*" information on the gender of the speaker, is shown. In this case, a single SVM has been trained with both male and female speeches. In more detail, the SVM has been trained and tested, considering the overall BES signals, by the k-fold cross-validation approach. The original BES signals are randomly partitioned into k equal size subsets. Among the k subsets, a single subset is retained to test the SVM, and the other $k - 1$ subsets are employed to train it. The cross-validation process is then repeated k times, with each of the k subsets used once as validation set. The obtained k results are then averaged to produce a single result. In this paper, in all considered cases, $k = 10$ has been employed.

TABLE 2. Confusion Matrix without applying any Gender Recognition.

	AN	BO	DI	FE	HA	NE	SA
AN	0.921	0	0.016	0.008	0.055	0	0
BO	0	0.852	0	0.025	0	0.123	0
DI	0.065	0.022	0.673	0.087	0.087	0.043	0.022
FE	0.029	0.043	0.043	0.783	0.043	0.043	0.014
HA	0.268	0.014	0.028	0.084	0.592	0.014	0
NE	0	0.240	0.013	0.025	0	0.709	0.013
SA	0	0.016	0	0.032	0	0.065	0.887

The obtained results are reported in Table 2, which reports the confusion matrix of the recognition rate. The reported values have been obtained by employing all the aforementioned 182 features (i.e., no PCA has been applied). In more detail, as for all the confusion matrices reported in this Performance Evaluation Section, the first row represents the recognized emotion while the first column contains the ground truth. For example, in Table 2, given anger (AN) as ground truth, the system "decides": AN in the 92.1% of the tests, never BO, DI in the 1.6% of the tests, and so on. Moreover, the mean value of the main diagonal of the matrix gives the average accuracy. If the mean of the main diagonal is computed from Table 2, it is possible to see that the method provides a good percentage of correct recognition for each emotion: the average value is about 77.4%. It is also clear that some emotions are better recognized (i.e., Anger and Sadness, recognized in the 92.1% and 88.7% of the cases, respectively) than other ones (such as Happiness, identified only in the 59.2% of the cases).

B. WITH GENDER RECOGNITION

Differently from the previous Section, now we evaluate the system performance when the "*a priori*" information on the gender of the speaker is used. This information has

been obtained by exploiting, in the testing phase, the Gender Recognition subsystem introduced in Section III-A and providing 100% gender recognition. In this case, as depicted in Fig. 3 and as extensively explained before, two SVMs, one for each gender, have been trained: the first SVM through male speeches signals, the second through female ones.

Also in this case, SVM training and testing phases have been carried out by two k-fold ($k = 10$) cross-validations and, again, the overall BES signals have been employed by dividing male speech from female speech signals.

Reported results show that the employment of information related to speaker gender allows improving the performance. The overall set of features (182) has been employed for these tests. In more detail, Table 3 and Table 4 show the confusion matrices concerning male and female speech signals, respectively.

TABLE 3. Confusion Matrix of Male Speech Signals by applying Gender Recognition.

	AN	BO	DI	FE	HA	NE	SA
AN	0.983	0	0	0.017	0	0	0
BO	0	0.743	0	0	0	0.200	0.057
DI	0.091	0.091	0.454	0.273	0	0	0.091
FE	0	0	0.056	0.805	0.056	0.056	0.028
HA	0.259	0	0	0.074	0.667	0	0
NE	0	0.103	0	0.026	0.026	0.846	0
SA	0	0.040	0	0	0	0	0.960

TABLE 4. Confusion Matrix of Female Speech Signals by applying Gender Recognition.

	AN	BO	DI	FE	HA	NE	SA
AN	0.881	0	0.030	0	0.089	0	0
BO	0	0.848	0	0	0	0.152	0
DI	0.029	0	0.857	0.057	0.029	0.029	0
FE	0.030	0	0	0.909	0.030	0.030	0
HA	0.341	0	0.091	0.045	0.477	0.045	0
NE	0	0.150	0	0	0	0.850	0
SA	0	0	0	0	0	0	1.000

The average percentage of correct classification if male speakers are recognized, is almost 79%. In the case of female speeches, the average value of the correct recognition is almost 84%. Globally, the performance in terms of correct emotion recognition (accuracy), in case of gender recognition is 81.5%, which represents a performance improvement of about the 5.3% with respect to the results shown in Table 2, where no gender information is employed.

C. FEATURES REDUCTION

As described previously in the paper, the considered approaches (i.e., a single SVM trained without distinguishing the gender and two SVMs driven by the proposed GR subsystem) can be implemented by employing a reduced number of features. A reduced number of features implies a reduction of the computational load needed to carry out the emotion recognition process and this opens the doors to the practical implementation of the proposed solution over mobile devices such as modern smartphones but performance must be satisfactory. In this paragraph a performance study of the accuracy by reducing the number of employed features, obtained by the PCA approach, is shown. In particular, as reported in Fig. 4, which shows the recognition percentage of the ER block by varying the number of used features, the average emotion recognition accuracy increases as the number of the employed features increases. The features are added in the emotion recognition process by following their *contribution* value, as described in Section III-B.2. From a numerical viewpoint, the accuracy reaches a values slightly above 70% of correctly classified emotions, if the GR subsystems is not employed, and a value of about 75%, if the GR subsystem drives the emotion recognition process. This performance is obtained when only 55 features are employed in the whole recognition process. The obtained accuracy is not so much lower than the accuracy (81.5%) obtained by using 182 features as reported above. We think that this performance decrease represents a reasonable trade-off between performance and computational load in view of future implementation over mobile platforms.

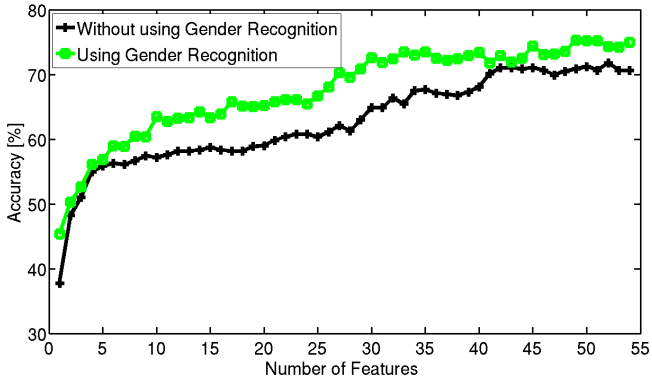

FIGURE 4. Recognition Percentage of the ER system versus the number of selected features.

D. SINGLE EMOTION DETECTION

In some applicative scenarios the recognition of a specific emotion with respect to the others is of great interest. For example, Safety and Security applications, which must recognize dangerous events in critical areas, such as train stations and airports, can exploit the recognition of fear detected by several smartphones users in the same zone to automatically monitor the whole area. Another possible example may concern Entertainment applications aimed at monitoring the positive opinion about plays, movies, concerts and shows: in all these cases the recognition of happiness among the other emotions can be a useful feedback.

For this reason, the proposed approach, which employs the gender information, has been compared, in terms of accuracy, with the traditional approach, which does not employ such information, to discriminate a particular emotion among

the others. Figures 5, 6, 7 and 8 show the performance, in terms of correct recognition of an emotion among the others for sadness, neutral state, happiness and fear, respectively. If the Gender Recognition subsystem is introduced, some emotional states such as sadness or neutral (Fig. 5 and 6) have a significant performance improvement. Concerning other emotions there is a limited advantage to use the proposed gender pre-processing. The recognition rate can slightly improve as in the happiness case (Fig. 7) but can also experience a slight deterioration as in the case of fear (Fig. 8).

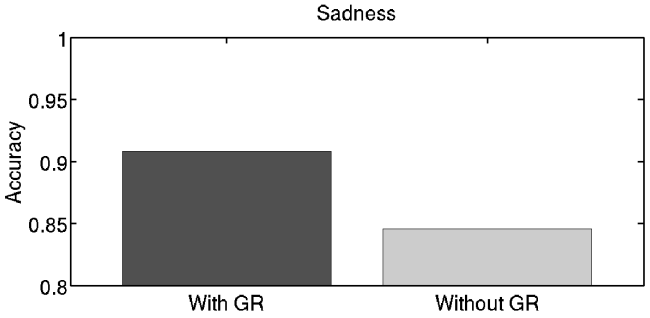

FIGURE 5. Sadness Recognition Rate of the ER system by using or not the Gender Recognition subsystem.

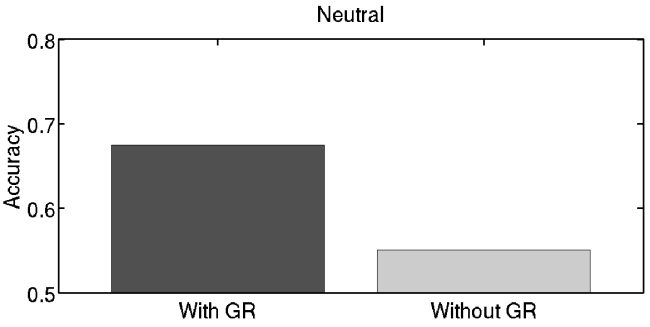

FIGURE 6. Neutral State Recognition Rate of the ER system by using or not the Gender Recognition subsystem.

FIGURE 7. Happiness Recognition Rate of the ER system by using or not the Gender Recognition subsystem.

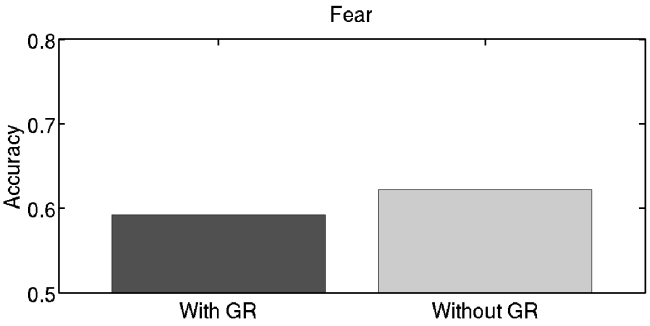

FIGURE 8. Fear Recognition Rate of the ER system by using or not the Gender Recognition subsystem.

V. CONCLUSION

The proposed system, able to recognize the emotional state of a person starting from audio signals registrations, is composed of two functional blocks: Gender Recognition (GR) and Emotion Recognition (ER). The former has been implemented by a Pitch Frequency Estimation method, the latter by two Support Vector Machine (SVM) classifiers (fed by properly selected audio features), which exploit the GR subsystem output.

The performance analysis shows the accuracy obtained with the adopted emotion recognition system in terms of recognition rate and the percentage of correctly recognized emotional contents. The experimental results highlight that the Gender Recognition (GR) subsystem allows increasing the overall emotion recognition accuracy from 77.4% to 81.5% due to the *a priori* knowledge of the speaker gender.

The results show that with the employment of a features selection algorithm, a satisfying recognition rate level can still be obtained also reducing the employed features and, as a consequence, the number of operations required to identify the emotional contents. This makes feasible future development of the proposed solution over mobile devices.

The obtained results underline that our system can be reliably used to identify a single emotion, or emotion category, versus all the other possible ones.

Possible future developments of this work can follow different directions: *i)* evaluation of the system performance by grouping the considered emotions in bigger sets (i.e., negative vs positive emotions); *ii)* evaluation of different classification algorithms; *iii)* implementation and related performance investigation of the proposed system on mobile devices; *iv)* computational load and energy consumption analysis of the implemented system.

ACKNOWLEDGMENT

The authors would like to thank Dr. Angelo Cirigliano and Dr. Fabio Patrone for their precious support in the testing phase of this research work and for their important suggestions.

REFERENCES

[1] F. Burkhardt, M. van Ballegooy, R. Englert, and R. Huber, "An emotion-aware voice portal," in *Proc. ESSP*, 2005, pp. 123–131.
[2] J. Luo, *Affective Computing and Intelligent Interaction*, vol. 137. New York, NY, USA: Springer-Verlag, 2012.
[3] R. W. Picard, *Affective Computing*. Cambridge, MA, USA: MIT Press, 2000.
[4] G. Chittaranjan, J. Blom, and D. Gatica-Perez, "Who's who with big-five: Analyzing and classifying personality traits with smartphones," in *Proc. 15th Annu. ISWC*, Jun. 2011, pp. 29–36.
[5] A. Vinciarelli, M. Pantic, and H. Bourlard. (2009, Nov.). "Social signal processing: Survey of an emerging domain," *Image Vision Comput.* [Online]. 27(12), pp. 1743–1759. Available: http://dx.doi.org/10.1016/j.imavis.2008.11.007
[6] Z. Zeng, M. Pantic, G. Roisman, and T. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009.
[7] A. Gluhak, M. Presser, L. Zhu, S. Esfandiyari, and S. Kupschick, "Towards mood based mobile services and applications," in *Proc. 2nd Eur. Conf. Smart Sens. Context*, 2007, pp. 159–174.
[8] K. K. Rachuri, M. Musolesi, C. Mascolo, P. J. Rentfrow, C. Longworth, and A. Aucinas, "Emotionsense: A mobile phones based adaptive platform for experimental social psychology research," in *Proc. UbiComp*, 2010, pp. 281–290.
[9] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognit.*, vol. 44, no. 3, pp. 572–587, 2011.
[10] R. Fagundes, A. Martins, F. Comparsi de Castro, and M. Felippetto de Castro, "Automatic gender identification by speech signal using eigenfiltering based on Hebbian learning," in *Proc. 7th Brazilian SBRN*, 2002, pp. 212–216.
[11] Y.-M. Zeng, Z.-Y. Wu, T. Falk, and W.-Y. Chan, "Robust GMM based gender classification using pitch and RASTA-PLP parameters of speech," in *Proc. Int. Conf. Mach. Learn. Cybern.*, 2006, pp. 3376–3379.
[12] Y.-L. Shue and M. Iseli, "The role of voice source measures on automatic gender classification," in *Proc. IEEE ICASSP*, Mar./Apr. 2008, pp. 4493–4496.
[13] F. Yingle, Y. Li, and T. Qinye, "Speaker gender identification based on combining linear and nonlinear features," in *Proc. 7th WCICA 2008*, pp. 6745–6749.
[14] H. Ting, Y. Yingchun, and W. Zhaohui, "Combining MFCC and pitch to enhance the performance of the gender recognition," in *Proc. 8th Int. Conf. Signal Process.*, vol. 1. 2006.
[15] D. Deepawale and R. Bachu, "Energy estimation between adjacent formant frequencies to identify speaker's gender," in *Proc. 5th Int. Conf. ITNG 2008*, pp. 772–776.
[16] F. Metze, J. Ajmera, R. Englert, U. Bub, F. Burkhardt, J. Stegmann, C. Muller, R. Huber, B. Andrassy, J. Bauer, and B. Littel, "Comparison of four approaches to age and gender recognition for telephone applications," in *Proc. IEEE ICASSP*, vol. 4. Apr. 2007, pp. IV-1089–IV-1092.
[17] E. Zwicker. (1961). "Subdivision of the audible frequency range into critical bands (frequenzgruppen)," *J. Acoust. Soc. Amer.* [Online]. 33(2), p. 248. Available: http://link.aip.org/link/?JAS/33/248/1
[18] R. Stibbard, "Automated extraction of ToBI annotation data from the reading/leeds emotional speech corpus," in *Proc. ISCA ITRW Speech Emotion*, Sep. 2000, pp. 60–65.
[19] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32–80, Jan. 2001.
[20] N. Campbell, "Building a corpus of natural speech-and tools for the processing of expressive speech-the JST CREST ESP project," in *Proc. 7th Eur. Conf. Speech Commun. Technol.*, 2001, pp. 1525–1528.
[21] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *Proc. Int. Joint Conf. Artif. Intell.*, vol. 14. 1995, pp. 1137–1145.
[22] W. H. Rogers and T. J. Wagner, "A fine sample distribution-free performance bound for local discrimination rules," *Ann. Stat.*, vol. 6, no. 3, pp. 506–514, 1978.
[23] R. Hanson, J. Stutz, and P. Cheeseman, *Bayesian Classification Theory*. Washington, DC, USA: NASA Ames Research Center, 1991.
[24] V. Vapnik, *The Nature of Statistical Learning Theory*. New York, NY, USA: Springer-Verlag, 1999.
[25] B. Yegnanarayana, *Artificial Neural Networks*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2004.
[26] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.
[27] A. A. Razak, R. Komiya, M. Izani, and Z. Abidin, "Comparison between fuzzy and NN method for speech emotion recognition," in *Proc. 3rd ICITA*, vol. 1. Jul. 2005, pp. 297–302.
[28] O. Pierre-Yves, "The production and recognition of emotions in speech: Features and algorithms," *Int. J. Human-Comput. Stud.*, vol. 59, no. 1, pp. 157–183, 2003.
[29] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.
[30] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proc. 4th ICSLP*, vol. 3. 1996, pp. 1970–1973.
[31] M. Westerdijk, D. Barber, and W. Wiegerinck, "Generative vector quantisation," in *Proc. 9th ICANN*, Jan. 1999, pp. 934–939.
[32] H. Harb and L. Chen, "Gender identification using a general audio classifier," in *Proc. ICME*, vol. 2. 2003, pp. II-733–II-736.
[33] M. Kotti and C. Kotropoulos, "Gender classification in two emotional speech databases," in *Proc. 19th ICPR 2008*, pp. 1–4.
[34] G. Tzanetakis, "Audio-based gender identification using bootstrapping," in *Proc. IEEE Pacific Rim Conf. Commun., Comput. Signal Process.*, Aug. 2005, pp. 432–433.
[35] D. Gerhard, "Pitch extraction and fundamental frequency: History and current techniques," Dept. Comput. Sci., Univ. Regina, Regina, SK, Canada, Tech. Rep., 2003.
[36] A. de Cheveigne and H. Kawahara. (2002). "YIN, a fundamental frequency estimator for speech and music," *J. Acoust. Soc. Amer.* [Online]. 111(4), pp. 1917–1930. Available: http://link.aip.org/link/?JAS/111/1917/1
[37] *Speech Signal Analysis*, Dept. Phonetics and Linguistics, London's Global Univ., London, U.K. [Online]. Available: http://www.phon.ucl.ac.uk/courses/spsci/matlab/lect10.html, Accessed: July 2013.
[38] J. Burg, *Modern Spectrum Analysis*, D. G. Childers, Ed. New York, NY, USA: IEEE Press, 1978, pp. 34–41.
[39] R. Snell and F. Milinazzo, "Formant location from LPC analysis data," *IEEE Trans. Speech Audio Process.*, vol. 1, no. 2, pp. 129–134, Apr. 1993.
[40] H. Abdi and L. J. Williams. (2010). "Principal component analysis," *Wiley Interdisciplinary Rev., Comput. Stat.* [Online]. 2(4), pp. 433–459. Available: http://dx.doi.org/10.1002/wics.101
[41] J. C. Platt, "Fast training of support vector machines using sequential minimal optimization," in *Advances in Kernel Methods*, B. Schölkopf, C. J. C. Burges, and A. J. Smola, Eds. Cambridge, MA, USA: MIT Press, 1999, pp. 185–208.
[42] F. Burkhardt, A. Paeschke, M. Rolfes, W. F. Sendlmeier, and B. Weiss, "A database of German emotional speech," in *Proc. Interspeech*, 2005, pp. 1517–1520.

IGOR BISIO (S'04–M'08) received the Laurea degree in telecommunication engineering and the Ph.D. degree from the University of Genoa, Genoa, Italy, in 2002 and 2006, respectively. He is currently an Assistant Professor and he is a member of the Digital Signal Processing and Satellite Communications and Networking Laboratories, University of Genoa. His main research activities concern resource allocation and management for satellite and space communication systems, and signal processing over smartphones.

ALESSANDRO DELFINO received the B.Sc. and M.Sc. degrees in telecommunication engineering from the University of Genoa, Genoa, Italy, in 2007 and 2010, respectively, with a thesis on audio fingerprinting. In 2010, he worked on MAC protocols for cognitive radio with the European funded Joint Research Center, Ispra, Italy. He is currently pursuing the Ph.D. degree with the University of Genoa. His main research activities concern audio fingerprinting and audio information retrieval, and cognitive radio.

FABIO LAVAGETTO is currently a Full Professor in telecommunications with the DITEN Department, University of Genoa, Genoa, Italy. Since 2008, he has been a Vice-Chancellor with responsibility for Research and Technology Transfer at the University of Genoa. Since 2005, he has been a Vice-Chair of the Institute for Advanced Studies in Information Technology and Communication. Since 1995, he has been the Head of research with the Digital Signal Processing Laboratory, University of Genoa. He was a General Chair of several international scientific conferences and has authored over 100 scientific publications in international journals and conferences.

MARIO MARCHESE (S'94–M'97–SM'04) received the Laurea degree (*cum laude*) in electronic engineering and the Ph.D. degree in telecommunications from the University of Genoa, Genoa, Italy, in 1992 and 1996, respectively. He is currently an Associate Professor with the DITEN Department, University of Genoa. His main research activities concern satellite and radio networks, transport layer over satellite and wireless networks, quality of service and data transport over heterogeneous networks, and applications for smartphones.

ANDREA SCIARRONE received the bachelor's and master's (*cum laude*) degrees from the University of Genoa, Genoa, Italy, in 2007 and 2009, respectively, both in telecommunication engineering, where he is currently pursuing the Ph.D. degree with the DITEN Department. His main research activities concern signal processing over portable devices, such as smartphones, context and location awareness, indoor localization, security and e-health applications, and android operative system.