Objective Human Affective Vocal Expression Detection and Automatic Classification with Stochastic Models and Learning Systems

V. Vieira, *Student Member, IEEE,* R. Coelho, *Senior Member, IEEE,* and F. M. de Assis

Abstract—**This paper presents a widespread analysis of affective vocal expression classification systems. In this study, state-of-the-art acoustic features are compared to two novel affective vocal prints for the detection of emotional states: the Hilbert-Huang-Hurst Coefficients (HHHC) and the vector of index of non-stationarity (INS). HHHC is here proposed as a nonlinear vocal source feature vector that represents the affective states according to their effects on the speech production mechanism. Emotional states are highlighted by the empirical mode decomposition (EMD) based method, which exploits the non-stationarity of the affective acoustic variations. Hurst coefficients (closely related to the excitation source) are then estimated from the decomposition process to compose the feature vector. Additionally, the INS vector is introduced as dynamic information to the HHHC feature. The proposed features are evaluated in speech emotion classification experiments with three databases in German and English languages. Three state-of-the-art acoustic features are adopted as baseline. The α-integrated Gaussian model (α-GMM) is also introduced for the emotion representation and classification. Its performance is compared to competing stochastic and machine learning classifiers. Results demonstrate that HHHC leads to significant classification improvement when compared to the baseline acoustic features. Moreover, results also show that α-GMM outperforms the competing classification methods. Finally, HHHC and INS are also evaluated as complementary features for the GeMAPS and eGeMAPS feature sets.**

Index Terms—**Hilbert-Huang transform, ensemble empirical mode decomposition, non-stationary degree, α-GMM, emotion classification.**

I. Introduction

AFFECTIVE states play an important role in the cognition, perception and communication of the human-being daily life. For instance, an unexpected event can motivate a happiness state. On the other hand, stressful situations may cause health problems. Automatic emotion recognition is especially important to improve communication between human and machine [1], [2]. In the literature, emotions are generally classified using physical or physiological signals such as speech [3], facial expression [4], and electrocardiogram (ECG)

V. Vieira is with the Post-Graduate Program in Electrical Engineering, Federal University of Campina Grande (UFCG), Campina Grande 58429-900, Brazil (e-mail: vinicius.vieira@ee.ufcg.edu.br).

R. Coelho is with the Laboratory of Acoustic Signal Processing (lasp.ime.eb.br), Military Institute of Engineering (IME), Rio de Janeiro 22290-270, Brazil (e-mail: coelho@ime.eb.br).

F. M. de Assis is with the Electrical Engineering Department, Federal University of Campina Grande (UFCG), Campina Grande 58429-900, Brazil (e-mail: fmarcos@dee.ufcg.edu.br).

[5]. Particularly, speech emotion recognition has received much research attention in the past few years [6]–[9]. In this scenario, many promising applications can be considered, such as security access, automatic translation, call-centers, mobile communication and human-robot interaction [10].

The speech production under emotions is affected by changes in muscle tension and in the breathing rate. These changes lead to different speech signals depending on the emotion. Figure 1 depicts amplitudes and corresponding spectrograms of speech signals produced with three affective expressions: Neutral, Anger, and Sadness. These signals were collected from the Berlin Database of Emotional Speech (EMO-DB) [11] and were spoken by the same female person and contain the same message. It can be noted that amplitudes and spectrograms are functions of the affective state.

In the context of social interactions, there is a large number of emotional states [12]. According to Ekman [2], there are certain emotions that can be naturally recognized by humans. Although this universality of the affective states discrimination, their decoding in the computational field is difficult. An *affective vocal print* is fundamental to a powerful recognition system. Thus, a key challenge is to define a feature that characterizes different emotions [3], [10]. In the literature, there is not yet a consensus about an effective acoustic feature for this task. In this sense, the choice of an attribute that shows meaningful information related to the physiological behavior of multiple affective states is a crucial search.

In [13], Teager-Energy-Operator (TEO) [14] based features were proposed for the classification of stress conditions. The idea was to capture nonlinear airflow structures of the acoustic signal induced by the speaker emotional state. Based on the fact that the excitation source signal reflects the speaker physiological behavior, vocal source features may also be applied for this purpose. Such features are less dependent on the linguistic content of speech [15], in comparison to spectral ones. In [8], the pH vocal source feature [16] was evaluated for emotion and stress classification. The authors showed that TEO features may be not suitable for emotion classification. Both pH and TEO features do not take into account the nonlinear effect of the speech production such as the non-stationarity of the affective acoustic variation and its dynamic behavior. These aspects are important to be exploited by an acoustic affective attribute.

One of the most common features applied as baseline in the literature and challenges is the mel-frequency cepstral coefficients (MFCC). This feature has been widely used for affective recognition due its success in other tasks, such as speech and speaker recognition [15], [17]. Nonetheless,
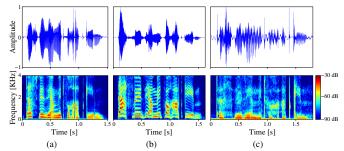
Fig. 1. Amplitudes and spectrograms of speech signals produced considering different emotional states: (a) Neutral, (b) Anger, and (c) Sadness.

other proposed features have shown superior performance than MFCC [8], [13], [17], [18]. For instance, the Hurst vector (pH) [16] achieves accuracy 6.8 percentage points (p.p.) higher than MFCC in emotions classification [8]. Some approaches have focused in recognition rates improvement, where several features are combined to form collections of low-level descriptors (LLDs) [10], [19]. This means that there is not yet a pure and established attribute for emotion classification. Furthermore, such studies are applied in the context of arousal and valence classification. Additionally, the scope of this present study is the representation of each affective state individually, which can improve the performance of classification tasks.

This work introduces a new nonlinear acoustic feature based on non-stationary effects of emotions. The empirical mode decomposition (EMD) [20] is applied to emphasize acoustic variations present in the speech signal. Hurst coefficients [21] are then estimated to characterize highlighted vocal source components. Finally, the Hilbert-Huang-Hurst Coefficients (HHHC) compose the affective vector on a frame-basis feature extraction. The combination of EMD with Hurst exponent is able to capture the non-stationary acoustic variations that occur during the speech production depending on the affective states. This aspect is still not well explored in the literature.

The index of non-stationarity (INS) [22] is here proposed as additional information to the HHHC feature vector. It dynamically describes the non-stationary behavior of affective speech samples. The α-GMM [23] is also introduced to classify emotional states. It is compared to classic Gaussian Mixture Models (GMM) [24] and Hidden Markov Models (HMM) [25] stochastic methods, and also machine learning approaches: Support Vector Machines (SVM) [26], Deep Neural Networks (DNN) [27], Convolutional Neural Networks (CNN) [28], and Convolutional Recurrent Neural Networks (CRNN) [29]). Experiments show the effectiveness of the new vocal source feature in different languages and scenarios. Several results demonstrate that HHHC is a 6-dimensional vector with robustness as a pure attribute for emotion. Additionally, HHHC contributes as complementary to GeMAPS and eGeMAPS [19] features sets to improve the classification rates.

This paper is organized as follows. Section II introduces the HHHC feature and presents the feature extraction procedure. The INS is also described in this section. The α-GMM and competing classifiers are presented in Section III. Evaluation experiments are described in Section IV and results are exhibited in Section V. Finally, Section VI concludes this work.

II. A New Nonlinear Acoustic Feature

The general idea of the Hilbert-Huang-Hurst Coefficients (HHHC) vector is to characterize the vocal source when affected by an emotional state. The affective content of the speech is highlighted by an adaptive method based on Hilbert-Huang transform (EMD). Instead of the original EMD, the ensemble EMD (EEMD) [30] is applied to analyze an improvement in the affective states detection. After the decomposition, Hurst coefficients, which are related to the excitation source, capture the nonlinear information from the emphasized acoustic variations. In [31], it was shown that acoustic sources have different degrees of non-stationarity. In this work, a vector of INS values is proposed to analyze and detect speech emotional states.

A. HHHC Feature

The HHHC vocal source feature is obtained by using the EMD-based approach and the estimation of Hurst coefficients from the decomposition process.

1) EMD/EEMD: EMD was introduced in [20] as a nonlinear time-domain adaptive method for decomposing non-stationary signals into a series of oscillatory modes. The general idea is to locally analyze a signal $x(t)$ between two consecutive extrema (minima or maxima). Then, two parts are defined: a local fast component, also called detail, $d(t)$, and the local trend or residual $a(t)$, such that $x(t) = d(t) + a(t)$. The detail function $d(t)$ corresponds to the first intrinsic mode function (IMF) and consists of the highest frequency component of $x(t)$. The subsequent IMFs are iteratively obtained from the residual of the previous IMF. The decomposition can be summarized in the following steps:

1) Identify all local extrema (minima and maxima) of $x(t)$;
2) Interpolate the local maxima and minima via cubic splines to obtain the upper ($e_{up}(t)$) and lower ($e_{lo}(t)$) envelopes, respectively;
3) Define the local trend as $a(t) = (e_{up}(t) + e_{lo}(t)) / 2$;
4) Calculate the detail component as $d(t) = x(t) - a(t)$.

Every IMF have zero mean, and the numbers of maxima and zero-crossings must be equal or differ by at most one. If the detail component $d(t)$ does not follow these properties, steps 1-4 are repeated with $d(t)$ in place of $x(t)$ until the new detail can be considered as an IMF. For the next IMF, the same procedure is applied on the residual $a(t) = x(t) - d(t)$.

Since an input signal $x(t)$ can be decomposed in a finite number of IMFs, the integrability property of the EMD can be expressed as $x(t) = \sum_{m=1}^{M} \mathrm{IMF}_m(t) + r(t)$, where $r(t)$ is the last residual sequence.

As an alternative for EMD, the EEMD method was proposed to avoid the *mode mixing* phenomena [30], which refer to IMF fluctuations that do not appear in the proper scale. Thus, the EEMD approach is expected to emphasize affective acoustic variations. Given the target signal $x(t)$, the EEMD method firstly generates an ensemble of I trials, $x^i(t)$, $i = 1, ..., I$, each consisting of $x(t)$ plus a white noise of finite amplitude, $w^i(t)$, i.e., $x^i(t) = x(t) + w^i(t)$. Each trial $x^i(t)$ is decomposed with EMD leading to M modes, $\mathrm{IMF}_m^i(t)$, $m = 1, ..., M$. Then, the m-th mode of $x(t)$ is obtained as the average of the I corresponding IMFs.
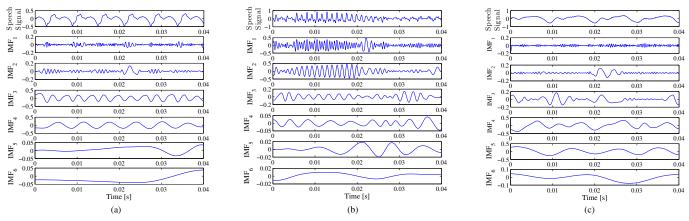
Fig. 2. First six IMFs obtained with EEMD from voiced speech segments: (a) Neutral, (b) Anger, and (c) Sadness.

Figure 2 shows the EEMD applied to three speech segments of 400 ms collected from EMO-DB [11]. Segments refer to Neutral speech (Figure 2a) and two basic emotions: Anger (Figure 2b) and Sadness (Figure 2c). The EEMD applies a high-frequency versus low-frequency separation between IMFs. Note that the affective signals have different non-stationary dynamic behaviors. For instance, IMFs 1 and 2 for Anger present amplitude values higher than for the other signals. On the other hand, the highest amplitude values are observed in the late three oscillations (IMFs 4, 5 and 6) of the Sadness state. This indicates that EEMD highlights the affective content of speech. For high-arousal emotions (e.g., Anger), non-stationary acoustic variations are more concentrated in the high-frequency IMFs, while the low-frequency ones capture the prevailing content from the low-arousal emotions (e.g., Sadness).

2) Hurst Coefficients: The Hurst exponent ($0 < H < 1$), or Hurst coefficient, expresses the time-dependence or scaling degree of a stochastic process [21]. Let a speech signal be represented by a stochastic process $x(t)$, with the normalized autocorrelation coefficient function $\rho(k)$, the H exponent is defined by the asymptotic behavior of $\rho(k)$ as $k \to \infty$, i.e., $\rho(k) \sim H(2H-1)k^{2(H-2)}$.

In this study, the H values are estimated from IMFs on a frame-by-frame basis using the wavelet-based estimator [32], which can be described in three main steps as follows:

1) Wavelet decomposition: the discrete wavelet transform (DWT) is applied to successively decompose the input sequence of samples into approximation ($a_w(j,n)$) and detail ($d_w(j,n)$) coefficients, where j is the decomposition scale ($j = 1, 2, ..., J$) and n is the coefficient index of each scale.

2) Variance estimation: for each scale j, the variance $\sigma^2 = (1/N_j)\sum_n d_w(j,n)^2$ is evaluated from detail coefficients, where N_j is the number of available coefficients for each scale j. In [32], it is shown that $E[\sigma_j^2] = C_H j^{2H-1}$, where C_H is a constant.

3) Hurst computation: a weighted linear regression is used to obtain the slope θ of the plot of $y_i = \log_2(\sigma_j^2)$ versus j. The Hurst exponent is estimated as $H = (1+\theta)/2$.

In [8], it was shown that H is related to the excitation source of emotional states. A high-arousal emotional signal has H values close to zero, while a low-arousal one has
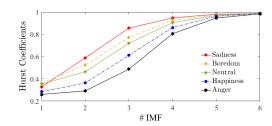

Fig. 3. Hurst mean values of six IMFs obtained from speech samples under five non-stationary emotional variations.
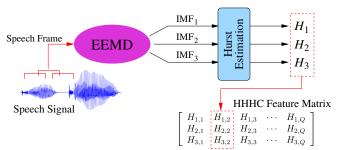

Fig. 4. An example of a HHHC vector extraction with 3 coefficients.

H values close to the unity. The authors extracted Hurst coefficients directly from the speech signal in a frame-basis for the pH feature [8]. In contrast, this present work deals with the estimation of Hurst values from IMFs of speech signals.

The HHHC vector for speech samples is illustrated in Figure 3. Signals are collected from the EMO-DB corresponding to five different emotional variations: Sadness, Boredom, Neutral, Happiness and Anger. A time duration of 40 s is considered for each emotional state. Six IMFs are obtained by the EEMD method, applied to speech segments of 80 ms and 50% overlapping. The Hurst exponent is computed and averaged from non-overlapping frames of 20 ms within each IMF, using Daubechies filters [33] with 12 coefficients and 3-12 scales in the wavelet-based Hurst estimator. It can be seen that the vocal source featured by Hurst coefficients are highlighted by the EEMD. Note that low-arousal emotions present the highest H values for the majority of the IMFs. For all the analyzed IMFs, high-arousal emotions have the lowest H averages.

3) HHHC Feature Extraction: The HHHC extraction of affective speech signals is performed in two main steps: signal decomposition using EMD or EEMD; and multi-channel estimation of the Hurst exponent. An example of the HHHC

vector estimation with 3 values of H is presented in Figure 4. The decomposition is applied to each segment of the input signal. The Hurst coefficients are obtained in a frame-by-frame basis from each IMF. Then, the feature matrix for HHHC is formed as an acoustic feature.

B. INS Vector

The INS is a time-frequency approach to objectively examine the non-stationarity of a signal [22]. The stationarity test is conducted by comparing spectral components of the signal to a set of stationary references, called *surrogates*. For this purpose, spectrograms of the signal and surrogates are obtained by means of the short time Fourier transform (STFT). Then, the Kullback-Leibler (KL) divergence is used to measure the distance between the spectrum of the analyzed signal and its global spectrum averaged over time. Given $\mathrm{KL}^{(x)}$ for the analyzed signal $x(t)$ and $\mathrm{KL}^{(s_j)}$ for the j surrogates obtained from $x(t)$. Since there are N short spectrograms, a variance measure, Θ, is obtained from the KL values:

$$
\begin{cases}
\Theta_0(j) = \mathrm{var}\left(\mathrm{KL}_n^{(s_j)}\right)_{n=1,...,N}, & j = 1, ..., J. \\
\Theta_1 = \mathrm{var}\left(\mathrm{KL}_n^{(x)}\right)_{n=1,...,N}.
\end{cases}
\tag{1}
$$

Finally, the INS is given by $\mathrm{INS} := \sqrt{\Theta_1/\langle\Theta_0(j)\rangle_j}$, where $\langle\cdot\rangle$ is the mean value of $\Theta_0(j)$. In [22], the authors considered that the distribution of the KL values can be approximated by a Gamma distribution. Therefore, for each window length T_h, a threshold γ can be defined for the stationarity test considering a confidence degree of 95%. Thus,

$$
\mathrm{INS} \begin{cases}
\leq \gamma & \text{, signal is stationary;} \\
> \gamma & \text{, signal is non-stationary.}
\end{cases}
\tag{2}
$$

Figure 5 depicts examples of the INS obtained from voiced segments of the Neutral state and two emotional variations: Anger and Sadness. The time scale T_h/T is the ratio between the length adopted in the short-time spectral analysis (T_h) and the total length ($T = 800$ ms) of the signal. Note that INS for both emotional states (red line) is higher than the threshold adopted in the test of non-stationarity (green line). However, the INS values vary from one emotional state to another. While the Neutral state has INS values in the range [50,100] for the majority of the observed time-scales, the INS for Sadness reaches a maximum value of 60. On the other hand, Anger presents INS greater than 100 for several time-scales.

III. CLASSIFICATION TASK

The α-integrated Gaussian Mixture Model is here proposed for acoustic emotion classification. The α-GMM was firstly proposed for speaker identification [23]. By introducing a factor of α, the modelling capacity of the GMM is extended, which is more suitable in acoustic variations conditions. The α-integration generalizes the linear combination adopted in the conventional GMM ($\alpha = -1$). For $\alpha < -1$, the α-GMM classifier emphasizes larger probability values and de-emphasizes smaller ones. Since affective states are assumed as acoustic variations added to speech in its production, it is understood that α-GMM increases the recognition performance. Similar to what was shown in [23], it was demonstrated
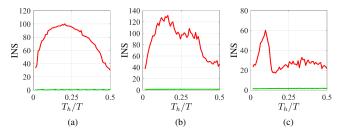

Fig. 5. INS computed from voiced segments considering emotional states: (a) Neutral, (b) Anger and (c) Sadness.

in [31] that α-GMM outperforms the conventional GMM. Hence, the HHHC is evaluated considering the α-GMM and the classical GMM ($\alpha = -1$). Five other classifiers are used for comparative evaluation purposes.

A. α-integrated Gaussian Mixture Model (α-GMM)

Given an affective state model λ_L, composed of M Gaussian densities $b_i(\mathbf{x})$, $i = 1, ..., M$, the α-integration of densities is defined as [23],

$$
p(\mathbf{x}|\lambda_L) = C \left[\sum_{i=1}^{M} \pi_i b_i(\mathbf{x})^{\frac{1-\alpha}{2}}\right]^{\frac{2}{1-\alpha}},
\tag{3}
$$

where π_i are non-negative mixture weights constrained to $\sum_{i=1}^{M} \pi_i = 1$, and C is a normalization constant. Note that $\alpha = -1$ corresponds to the conventional GMM.

Models λ_L are completely parametrized by mean vectors, covariance matrices, and weights of Gaussian densities. These parameters are estimated using an adapted expectation-maximization (EM) algorithm as to maximize the likelihood function $p(\mathbf{X}|\lambda_L) = \prod_{t=1}^{Q} p(\mathbf{x}_t|\lambda_L)$, where $\mathbf{X} = [\mathbf{x}_1\mathbf{x}_2\ldots\mathbf{x}_Q]$ is the feature matrix extracted from the training speech segment Φ_L of the affective state L.

B. Hidden Markov Models (HMM)

The HMM consists of finite internal states that generate a set of external events (observations). These states are hidden for the observer, and capture the temporal structure of an affective speech signal. Mathematically, the HMM can be characterized by three fundamental problems:

1) Likelihood: Given an HMM $\lambda_L = (A, B)$ with K states, and an observation sequence $\mathbf{x}$, determine the likelihood $p(\mathbf{x}|\lambda_L)$, where A is a matrix of transitions probabilities a_{jk}, $j, k = 1, 2, ..., K$, from state j to state k, and B is the set of densities b_j;
2) Decoding: Given an observation sequence $\mathbf{x}$ and an HMM λ_L, discover the sequence of hidden states;
3) Learning: Given an observation sequence $\mathbf{x}$ and the set of states in the HMM, learn the parameters A and B.

The standard algorithm for HMM training is the forward-backward, or Baum-Welch algorithm [34]. It obtains A and B matrices which maximizes the likelihood $p(\mathbf{x}|\lambda_L)$. The Viterbi algorithm is commonly used for decoding [35].

C. Support Vector Machines (SVM)

SVM [26] is a classical supervised machine learning model widely applied for data classification. The general idea is to

find the optimal separating hyperplane which maximizes the margin on the training data. For this purpose, it transforms input vectors into a high-dimensional feature space using a nonlinear transformation (with a kernel function). Given a training set $\{u_\xi\}_{\xi=1}^N = \{(\mathbf{x}_\xi, L_\xi)\}_{\xi=1}^N$, where $L_\xi \in \{-1, +1\}$ represents the affective state L of the utterance ξ. Thus, the classifier is a hyperplane defined as $g(\mathbf{x}) = \boldsymbol{w}^{\mathrm{T}}\mathbf{x} + b$, where $\boldsymbol{w}$ is the gradient vector which is perpendicular to the hyperplane, and b is the offset of the hyperplane from the origin. The side of the hyperplane which belongs the utterance can be indicated by $L_\xi g(\mathbf{x}_\xi)$. For $L_\xi = +1$, $L_\xi g(\mathbf{x}_\xi)$ must be greater than 1, while $L_\xi g(\mathbf{x}_\xi)$ is required to be smaller than -1 for $L_\xi = -1$. Then, the hyperplane is chosen by the solution of the optimization problem of minimizing $\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$ subject to $L_\xi\left(\boldsymbol{w}^{\mathrm{T}}\mathbf{x} + b\right) \geq 1, \xi = 1, 2, ..., N$.

In this work, the input data for the SVM classifier is obtained from mean vectors of feature matrices. This statistic was more prominent than others, such as median and maximum value, as observed in [36]. Radial Basis Function (RBF) is used as the SVM kernel.

D. Deep Neural Networks (DNN)

DNN is one of the most prominent methods for machine learning tasks such as speech recognition [37], separation [38], and emotion classification [9]. The deep learning concept can be applied for architectures such as feedfoward multilayer perceptrons (MLPs), convolutional neural networks (CNNs) and recurrent neural networks (RNNs) [39]. In this work, it is considered MLP that has feedforward connections from the input layer to the output layer, with sigmoid activation function y_j for the neuron j, $y_j = 1/(1 + e^{-x_j})$, where $x_j = b_j + \sum_i y_i w_{ij}$ is a weighted sum of the previous neurons with a bias b_j [37].

E. Convolutional Neural Networks (CNN)

Convolutional Neural Networks [28] have been widely adopted in the acoustic signal processing area, particularly for sound classification [40], [41] and sound event detection [42]. CNNs extend the multilayer perceptrons model by the introduction of a group of convolutional and pooling layers. The convolutional kernels are proposed to better capture and classify the spectro-temporal patterns of acoustic signals. Pooling operations are then applied for dimensionality reduction between convolutional layers.

F. Convolutional Recurrent Neural Networks (CRNN)

CRNNs [29] consist on the combination of CNNs with Recurrent Neural Networks (RNN). The idea is to improve the CNN by learning spectro-temporal information of relatively longer events that are not captured by the convolutional layers. For this purpose, recurrent layers are applied to the output of the convolutional layers to integrate the information of earlier time windows. In the literature, CNNs and RNNs have been successfully combined for music classification [43] and sound event detection [29]. In this work, a single feedforward layer with sigmoid activation function that follows the recurrent layers is considered as the output layer of the network [29].
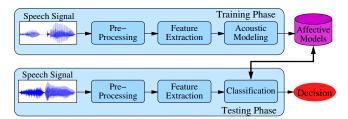

Fig. 6. Affective vocal expression: classification system diagram.

IV. EXPERIMENTAL SETUP

Extensive experiments are carried out to evaluate the proposed HHHC acoustic feature. Figure 6 illustrates the classification system used in the experiments. Affective models are generated in the training phase after pre-processing and feature extraction. During tests, for each voiced speech signal, the extracted feature vector is compared to each model. The leave-one-speaker-out (LOSO) methodology [7] is adopted to achieve speaker independence. For all databases, the modelling of each affective state is conducted with 32 s randomly selected from the training data. Test experiments are applied to 800 ms speech segments of each emotion of the testing speaker. The detection of emotional content in instances which last less than 1 s is suitable for real-life situations [10].

The α-GMM is performed with five values of α: -1 (classical GMM), -2, -4, -6 and -8. Affective models are composed of 32 Gaussian densities with diagonal covariance matrices. The HMM is implemented using the HTK toolkit [44] with the left-to-right topology. For each affective condition, it is used five HMM states with one single Gaussian mixture per state. The SVM implementation is carried out with the LIBSVM [45], using the "one-versus-one" strategy. The search for the optimal hyperplane is conducted in a grid-search procedure for the RBF kernel, with the controlling parameters being evaluated for $c \in (0, 10)$ and $\gamma \in (0, 1)$. The DNNs consider multilayer perceptrons with three hidden layers [38]. The networks are trained with the standard backpropagation algorithm with dropout regularization (dropout rate 0.2). It is not used any unsupervised pretraining. The momentum rate used is 0.5. Sigmoid activation functions are used in the output layer, while linear functions are used for the rest. CNNs and CRNNs are implemented with three convolutional layers followed by max pooling operation with (2,2,2) and (5,4,2) pool arrangements, respectively [29]. A single recurrent layer is used to compose the CRNN.

In order to verify the improvement in classification rates for emotion recognition, the proposed HHHC vector is experimented as complementary to collections of features such as GeMAPS [19]. For this purpose, binary arousal and valence classification is carried out by using the SVM classifier.

A. Speech Emotion Databases

Three databases are considered in the experiments: EMO-DB [11], IEMOCAP (Interactive Emotional Dyadic Motion Capture) [46], and SEMAINE (Sustained Emotionally colored Machine-human Interaction using Nonverbal Expression) [47]. Only the voiced segments of speech are considered in the experiments. For this purpose, the pre-processing step selects

TABLE I
ACCURACY RATES (%) OF 5 EMOTIONAL STATES WITH THE HHHC AND BASELINE FEATURES FOR EMO-DB.

Classifier	Actual Emotion	HHHC feature					HHHC + INS					pH feature					MFCC feature					TEO feature				
		Ang.	Hap.	Neu.	Bor.	Sad.	Ang.	Hap.	Neu.	Bor.	Sad.	Ang.	Hap.	Neu.	Bor.	Sad.	Ang.	Hap.	Neu.	Bor.	Sad.	Ang.	Hap.	Neu.	Bor.	Sad.
α-GMM Classifier	Anger	**86**	14	0	0	0	**88**	12	0	0	0	**82**	18	0	0	0	**80**	20	0	0	0	43	41	16	0	0
	Happiness	35	**65**	0	0	0	32	**68**	0	0	0	41	**55**	4	0	0	18	**80**	2	0	0	31	**55**	10	4	0
	Neutral	0	0	**86**	14	0	0	0	**87**	13	0	0	6	**69**	14	11	0	17	**55**	19	9	8	18	**47**	27	0
	Boredom	0	0	14	**71**	15	0	0	10	**77**	13	0	4	20	**43**	33	0	6	30	**35**	29	6	14	24	**43**	13
	Sadness	0	0	0	12	**88**	0	0	0	11	**89**	0	2	8	12	**78**	0	2	8	22	**68**	4	0	6	14	**76**
	Average:	79.2					81.8					65.4					63.6					52.8				
HMM Classifier	Anger	**76**	24	0	0	0	**77**	23	0	0	0	**78**	22	0	0	0	**74**	24	2	0	0	28	52	20	0	0
	Happiness	33	**67**	0	0	0	30	**70**	0	0	0	32	**64**	4	0	0	25	**70**	5	0	0	31	**59**	5	5	0
	Neutral	0	0	**81**	19	0	0	0	**84**	16	0	0	6	**64**	20	10	0	19	**48**	23	10	10	34	**24**	32	0
	Boredom	0	0	15	**68**	17	0	0	14	**71**	15	0	5	31	**33**	31	0	8	34	**28**	30	3	6	26	**51**	14
	Sadness	0	0	0	19	**81**	0	0	0	18	**82**	0	3	8	15	**74**	0	5	11	25	**59**	4	0	6	15	**75**
	Average:	74.6					76.8					62.6					55.8					47.4				
SVM Classifier	Anger	**72**	28	0	0	0	**73**	27	0	0	0	**69**	30	1	0	0	**63**	30	7	0	0	20	56	24	0	0
	Happiness	37	**63**	0	0	0	36	**64**	0	0	0	35	**57**	8	0	0	27	**65**	8	0	0	30	**55**	10	5	0
	Neutral	0	0	**64**	34	2	0	0	**67**	23	0	0	8	**56**	24	12	0	20	**43**	25	12	13	36	**20**	31	0
	Boredom	0	0	20	**51**	29	0	0	19	**52**	29	0	9	28	**27**	36	0	11	37	**19**	33	4	7	27	**47**	15
	Sadness	0	0	0	29	**71**	0	0	0	27	**73**	0	2	10	20	**68**	0	12	24	35	**29**	7	7	0	17	**69**
	Average:	64.2					65.8					55.4					43.8					42.2				

frames of 16 ms with high energy and low zero crossing rate. The sampling rate used for all databases is 8 kHz.

EMO-DB consists of ten actors (5 women and 5 men) that uttered ten sentences in German with archetypical emotions. In this work, five emotional states are considered: Anger, Happiness, Neutral, Boredom and Sadness. Although EMO-DB comprises seven emotions (including disgust and fear), the experiments with five of them are carried out in order to show the power of an acoustic feature in characterize emotions that are naturally recognized by humans. Thus, five emotions were chosen to show the effectiveness of the HHHC vector. The entire set of voiced speech samples for each emotional state has 40 s.

IEMOCAP is composed of conversations of both scripted and spontaneous scenarios in English language. Ten actors (5 women and 5 men) were recorded in dyadic sessions in order to facilitate a more natural interaction of the targeted emotion. Since it is analyzed short emotional instances in the test phase, it is used a portion of the IEMOCAP database, although it comprises 12 hours of recordings. It is considered four emotional states: Anger, Happiness, Neutral and Sadness. A total of 10 minutes of voiced content from each emotional state is used in the experiments, where it is considered 5 minutes of both tasks (scripted and spontaneous scenarios).

The SEMAINE database features 150 participants (undergraduate and postgraduate students from eight different countries). The Sensitive Artificial Listener (SAL) scenario was used in conversations in English. Interactions involve a "user" (human) and an "operator" (either a machine or a person simulating a machine). In this work, it is considered recordings from ten participants (5 women and 5 men). From 27 categories (styles), 4 emotional states were selected: Anger, Happiness, Amusement and Sadness. The set of voiced speech samples for each emotional state has 90 s.

B. Extracted Features

6-dimensional HHHC vectors are extracted according to the procedure presented in the Section II-A. In the EEMD-based analysis, it is experimented 11 Gaussian noise levels, considering the noise standard deviation (std) in the range $[0.005, 0.1]$. The robustness of the HHHC is also verified using the INS in the feature vector (HHHC+INS). For each IMF, the INS values are computed with ten different observation scales, $T_h/T \in [0.0015, 0.5]$.

For the performance comparison and feature fusion, MFCC, TEO-CB-Auto-Env and pH vector are used in the experiments. 12-dimensional MFCC vectors are obtained from speech frames of 25 ms, with a frame rate of 10 ms. For the TEO-CB-Auto-Env (TEO feature), vectors with 16 coefficients are extracted from 75 ms speech samples, with 50% overlapping. The estimation of the pH feature is conducted in frames of 50 ms, every 10 ms, using the Daubechies wavelet filters with 12 coefficients (2-12 scales). Fusion procedures are carried out for an improvement provided by the proposed HHHC in the recognition rates of the baseline features.

V. RESULTS

This Section presents accuracies results obtained in speech emotion classification. For this purpose, confusion matrices are obtained considering the proposed HHHC and baseline features. Tables I, II and III present accuracies achieved for the EMO-DB, IEMOCAP and SEMAINE databases, respectively. They show confusion matrices obtained with α-GMM, HMM and SVM classifiers for the HHHC, HHHC+INS, and baseline features. The EMD-based HHHC already outperforms competing attributes. However, the EEMD-based approach reaches superior accuracies. Results for HHHC are achieved with the EEMD-based approach considering low Gaussian noise level (0.005≤ std ≤0.02).

A. Results with EMO-DB

For the α-GMM, the proposed HHHC feature achieves the best average accuracy (79.2%) with three values of α (−4, −6 and −8). This value is greater than the average accuracy

TABLE II
ACCURACY RATES (%) OF 4 EMOTIONAL STATES WITH THE HHHC AND BASELINE FEATURES FOR IEMOCAP.

	Actual Emotion	HHHC feature				HHHC + INS				pH feature				MFCC feature				TEO feature			
		Classified Emotion				Classified Emotion				Classified Emotion				Classified Emotion				Classified Emotion			
		Ang.	Hap.	Neu.	Sad.	Ang.	Hap.	Neu.	Sad.	Ang.	Hap.	Neu.	Sad.	Ang.	Hap.	Neu.	Sad.	Ang.	Hap.	Neu.	Sad.
α-GMM Classifier	Anger	**66**	23	9	2	**68**	23	9	0	**59**	24	13	4	**59**	16	15	10	**40**	25	24	11
	Happiness	26	**55**	15	4	26	**57**	15	2	28	**47**	17	8	28	**43**	20	9	33	**36**	21	10
	Neutral	10	12	**61**	17	9	11	**63**	17	12	15	**52**	21	16	11	**47**	26	7	24	**37**	32
	Sadness	7	9	22	**62**	6	9	22	**63**	9	13	25	**53**	9	11	26	**54**	8	5	23	**64**
		Average: **61.0**				Average: **62.8**				Average: **52.8**				Average: **50.8**				Average: **44.2**			
HMM Classifier	Anger	**55**	28	12	5	**58**	28	13	1	**57**	26	13	4	**50**	19	18	13	**37**	26	25	12
	Happiness	31	**45**	19	5	30	**48**	18	4	33	**42**	17	8	30	**37**	22	11	35	**31**	22	12
	Neutral	10	15	**54**	21	11	13	**57**	19	12	15	**49**	24	16	12	**44**	28	8	25	**33**	34
	Sadness	7	12	27	**54**	6	10	26	**58**	10	14	27	**49**	10	12	28	**50**	9	8	24	**59**
		Average: **52.0**				Average: **55.3**				Average: **49.3**				Average: **45.3**				Average: **40.0**			
SVM Classifier	Anger	**49**	31	14	6	**51**	31	14	4	**49**	30	15	6	**40**	22	23	15	**27**	30	29	14
	Happiness	30	**35**	28	7	30	**38**	27	5	29	**30**	26	15	32	**32**	24	12	37	**25**	24	14
	Neutral	15	20	**39**	26	15	19	**40**	26	17	24	**32**	27	18	15	**31**	36	9	27	**26**	38
	Sadness	7	14	33	**46**	7	14	32	**47**	12	15	33	**40**	13	15	31	**41**	9	9	27	**55**
		Average: **42.3**				Average: **44.0**				Average: **37.8**				Average: **36.0**				Average: **33.3**			

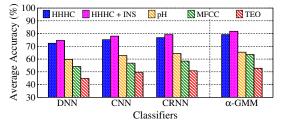Fig. 7. Average accuracies of EMO-DB obtained with α-GMM and Neural Network classifiers.

Fig. 9. Average accuracies of IEMOCAP obtained with α-GMM and Neural Network classifiers.

Fig. 8. Classification accuracies with feature fusion and α-GMM classifier of emotional states from EMO-DB.
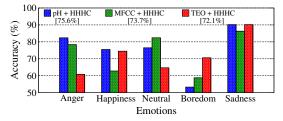

Fig. 10. Classification accuracies with feature fusion and α-GMM classifier of emotional states from IEMOCAP.

achieved with pH for $\alpha = -2$ (65.4%). The HHHC also outperforms in 15.6 p.p. the average accuracy of MFCC (63.6%), and reaches 26.4 p.p. over the TEO feature (52.8%). The INS information contributes for an increasing of more than 2 p.p. over the HHHC. The HHHC enables almost 60.0% of recognition for each considered emotional state using α-GMM. For all considered feature sets, the α-GMM (including the original GMM) outperforms the HMM and SVM classifiers.

Figure 7 presents the average classification accuracies obtained with the proposed and baseline features considering the Neural Network classifiers. Average results obtained with the α-GMM are also shown in Figure 7. Note that HHHC and HHHC+INS achieve the best results for all classifiers. For the CRNN, which outperforms DNN and CNN, HHHC leads to an improvement of 12.4 p.p. over pH: from 64.4% to 76.8%. For this classifier, the average accuracy obtained with HHHC+INS achieves 79.2%, i.e., 2.4 p.p. higher than HHHC. It can also be noticed that the introduced α-GMM achieves the best classification accuracies for all features sets.
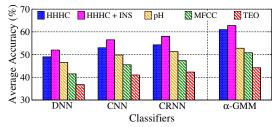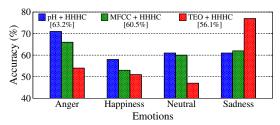
For HHHC+INS features, for example, the average accuracy with α-GMM is 2.6 p.p. greater than CRNN.

Figure 8 shows the identification accuracy with α-GMM for the feature fusion between HHHC and competing features. The best average accuracy attained with the pH+HHHC fusion (75.6% with $\alpha = -6$) is 10.2 p.p. higher than that achieved with pH only (65.4%). The MFCC+HHHC fusion reaches the best accuracy (73.7%) with $\alpha = -8$. It means that HHHC increases in almost 10 p.p. the recognition rate provided by the MFCC feature. About the TEO+HHHC fusion, the best average accuracy is 72.1% with $\alpha = -6$ and $\alpha = -8$. This means an improvement of 19.2 p.p. provided by the HHHC for the TEO-based feature.

B. Results with IEMOCAP

It can be seen from Table II that, for all considered feature sets, the α-GMM achieves superior accuracies over the HMM and the SVM classifiers. Only HHHC and HHHC+INS reach average accuracies over 60.0%. These values are achieved

TABLE III
ACCURACY RATES (%) OF 4 EMOTIONAL STATES WITH THE HHHC AND BASELINE FEATURES FOR SEMAINE.

	Actual Emotion	HHHC feature				HHHC + INS				pH feature				MFCC feature				TEO feature			
		Classified Emotion				Classified Emotion				Classified Emotion				Classified Emotion				Classified Emotion			
		Ang.	Hap.	Amu.	Sad.	Ang.	Hap.	Amu.	Sad.	Ang.	Hap.	Amu.	Sad.	Ang.	Hap.	Amu.	Sad.	Ang.	Hap.	Amu.	Sad.
α-GMM Classifier	Anger	**50**	23	20	7	**51**	23	20	6	**50**	22	20	8	**42**	29	16	13	**34**	24	22	20
	Happiness	14	**57**	25	4	14	**59**	25	2	17	**51**	27	5	18	**52**	26	4	29	**33**	29	9
	Amusement	14	26	**51**	9	13	24	**55**	8	16	26	**48**	10	15	30	**47**	8	19	25	**35**	21
	Sadness	6	15	19	**60**	5	15	17	**63**	8	15	23	**54**	9	11	25	**55**	3	16	20	**61**
		Average: **54.5**				Average: **57.0**				Average: **50.8**				Average: **49.0**				Average: **40.8**			
HMM Classifier	Anger	**45**	26	22	7	**46**	25	22	7	**45**	25	22	8	**38**	31	17	14	**28**	26	24	22
	Happiness	17	**50**	28	5	17	**53**	28	2	19	**47**	29	5	19	**49**	28	4	30	**31**	30	9
	Amusement	14	29	**48**	9	13	27	**51**	9	16	28	**45**	11	16	31	**42**	11	20	27	**31**	22
	Sadness	8	18	22	**52**	5	18	22	**55**	8	18	27	**47**	10	13	30	**47**	3	18	24	**55**
		Average: **48.8**				Average: **51.3**				Average: **46.0**				Average: **44.0**				Average: **36.2**			
SVM Classifier	Anger	**39**	28	24	9	**41**	28	24	7	**38**	29	25	8	**30**	34	20	16	**18**	30	28	24
	Happiness	20	**43**	32	5	19	**45**	31	5	22	**40**	33	5	21	**41**	33	5	33	**22**	35	10
	Amusement	16	32	**43**	9	15	30	**44**	11	18	30	**39**	13	18	34	**35**	13	21	29	**24**	26
	Sadness	9	20	25	**46**	7	20	26	**47**	9	20	31	**40**	11	15	35	**39**	3	21	29	**47**
		Average: **42.8**				Average: **44.3**				Average: **39.3**				Average: **36.3**				Average: **27.8**			

Fig. 11. Average accuracies of SEMAINE obtained with α-GMM and Neural Network classifiers.

Fig. 12. Classification accuracies with feature fusion and α-GMM classifier of emotional states from SEMAINE.

using the α-GMM with $\alpha = -8$. In comparison to baseline features, HHHC obtained an average accuracy 8 p.p. over the pH vector ($\alpha = -8$), 10 p.p. over the MFCC ($\alpha = -4$) and 15 p.p over the TEO-based feature ($\alpha = -6$). For each considered emotional state, the α-GMM approach achieves more than 50.0% accuracies with HHHC. Furthermore, α-GMM provides an improved performance with baseline features, in comparison to HMM and SVM approaches.

Figure 9 presents the average classification accuracies of IEMOCAP considering α-GMM and Neural Network classifiers. As in the EMO-DB, HHHC outperforms the pH, MFCC and TEO features for all classifiers. For the CRNN, HHHC achieves an average accuracy of 54.3%, which is 3.0 p.p., 7.0 p.p., and 12.0 p.p. greater than pH, MFCC, and TEO, respectively. Moreover, HHHC+INS leads to the best results for all scenarios. The α-GMM also outperforms the competing classifiers for all features sets.

Figure 10 depicts results of the feature fusion using the α-GMM for the HHHC and baseline features in the IEMO-CAP database. The pH+HHHC fusion achieves an accuracy of 63.2% ($\alpha = -8$), which outperforms both pH (52.8%) and HHHC+INS (62.8%). The fusion of Hurst-based features (pH+HHHC) indicates that the relation between H and the excitation source enables a high performance in the separation of basic emotions. As for the MFCC+HHHC fusion, HHHC leads to the MFCC an improvement in the average accuracy from 50.8% to 60.5% ($\alpha = -4$). Considering the TEO+HHHC fusion, the best average accuracy (56.1%) is achieved with $\alpha = -4$, which is 11.9 p.p. higher than that obtained with the TEO-based feature only.
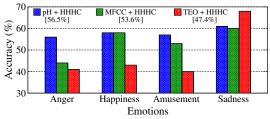
C. Results with SEMAINE

The best average accuracy is achieved with HHHC and HHHC+INS (refer to Table III): 54.5% and 57.0%, respectively, using α-GMM with $\alpha = -6$. These results are greater than 50.8% for pH ($\alpha = -4$), 49.0% for the MFCC ($\alpha = -6$), and 40.8% for TEO-based feature ($\alpha = -8$). An important issue on the SEMAINE database is mainly concerned to the Happiness and Amusement states recognition. Although these emotions present similar behavior, the HHHC shows to be able to recognize both of them with an accuracy over 50.0% in the classification provided by the α-GMM. For baseline features, the α-GMM reaches more than 4 p.p. over HMM and 10 p.p. over SVM. The α-GMM outperforms HMM and SVM for all considered emotional states. According to the average classification results shown in Figure 11, α-GMM also outperforms the competing DNN, CNN and CRNN classifiers. For these classifiers, HHHC and HHHC+INS also achieve the best average results.

The best recognition rates on the feature fusion task with the HHHC and the baseline features using α-GMM are shown in Figure 12. The pH+HHHC fusion attains an average accuracy of 56.5%, which represents an improvement over pH and HHHC features. With the MFCC+HHHC feature fusion, it is observed an enhancement from 49.0% to 53.6% in the recognition rate, with $\alpha = -6$. The HHHC provides an improvement of more than 6 p.p. when compared to the TEO-based feature (47.4%, $\alpha = -8$). The proposed feature is also very promising for discriminant learning strategies [9] applied to DNN and Deep Convolutional Neural Networks (DCNN) methods for speech emotion classification.

TABLE IV
CLASSIFICATION OF BINARY AROUSAL AND VALENCE FOR EMO-DB.

Feature Set	UAR (%) with SVM	
	Arousal	Valence
HHHC	80.5	67.8
HHHC+INS	83.2	69.9
GeMAPS	93.2	74.4
eGeMAPS	93.9	74.8
GeMAPS+HHHC	96.1	79.1
GeMAPS+HHHC+INS	**97.6**	**80.4**
eGeMAPS+HHHC	96.7	81.3
eGeMAPS+HHHC+INS	**98.4**	**82.1**

D. HHHC Complementarity Aspect

In order to evaluate the complementarity of the HHHC feature vector to collections of features sets, binary arousal and valence emotion classification are carried out considering all emotions of EMO-DB. The GeMAPS feature set and its extended version (eGeMAPS) [19] are adopted for this purpose. The experimental setup is similar to [19] with LOSO cross-validation with eight folds, where the speaker IDs are randomly arranged into eight speaker groups. The SVM method is applied for the classification procedure with the LIBSVM toolkit and the same parameters presented in Section IV. Table IV shows results of UAR (Unweighted Average Recall) obtained from experiments with GeMAPS, eGeMAPS, HHHC, HHHC+INS, and the feature fusion of the proposed acoustic feature with the comparative feature sets. Note that, for arousal evaluation, GeMAPS and eGeMAPS reach more than 93% UAR while HHHC and HHHC+INS achieve 80.5% and 83.2%, respectively. While the standard feature sets needs 62 and 88 features (GeMAPS and eGeMAPS, respectively) for this result, HHHC shows interesting accuracy for a low dimensional feature. However, HHHC and HHHC+INS contribute for an improvement in the UAR obtained with GeMAPS and eGeMAPS. For instance, eGeMAPS+HHHC+INS reaches 98.4% UAR. In valence classification, HHHC and HHHC+INS also contribute to the feature sets. GeMAPS performance is improved from 74.4% to 80.4% with HHHC+INS, while eGeMAPS reaches 82.1% with this fusion. This experiment demonstrates the complementarity potential of the HHHC to the GeMAPS and eGeMAPS features sets.

VI. CONCLUSION

This work introduced the HHHC nonlinear vocal source feature vector for speech emotion classification. The INS was used as dynamic information for the HHHC vector. Furthermore, the α-GMM approach was proposed for this classification task. It was compared to HMM, SVM, DNN, CNN, and CRNN. The best average classification accuracies were obtained using the α-GMM. In comparison to baseline features, HHHC obtained superior accuracy considering three different databases. On the feature fusion, HHHC provides an improved performance for all considered baseline features. As for the EMO-DB, the highest classification accuracy was 81.8% with HHHC+INS. For the IEMOCAP database, it was reached an average accuracy of 63.2% with pH+HHHC. In the SEMAINE context, the best average accuracy was 57.0% with HHHC+INS. The superior performance of the proposed

feature showed that the HHHC is very promising for affective state representation and for classification tasks. Also, the HHHC complementarity to GeMAPS features set was verified by the improvement in the recognition rates in binary arousal and valence emotion classification.

REFERENCES

[1] 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 Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[2] P. Ekman, *The Handbook of Cognition and Emotion*. Wiley Online Library, 1999, ch. Basic Emotions, pp. 45–60.

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

[4] E. Barakova, R. Gorbunov, and M. Rauterberg, "Automatic interpretation of affective facial expressions in the context of interpersonal interaction," *IEEE Transactions on Human-Machine Systems*, vol. 45, no. 4, pp. 409–418, Aug 2015.

[5] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, "ECG pattern analysis for emotion detection," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 102–115, January 2012.

[6] A. Tawari and M. M. Trivedi, "Speech emotion analysis: Exploring the role of context," *IEEE Transactions on Multimedia*, vol. 12, no. 6, pp. 502–509, October 2010.

[7] B. Schuller, B. Vlasenko, F. Eyben, G. Rigoll, and A. Wendemuth, "Acoustic emotion recognition: A benchmark comparison of performances," in *IEEE Workshop on Automatic Speech Recognition & Understanding*, 2009, pp. 552–557.

[8] L. Zão, D. Cavalcante, and R. Coelho, "Time-frequency feature and AMS-GMM mask for acoustic emotion classification," *IEEE Signal Processing Letters*, vol. 21, no. 5, pp. 620–624, May 2014.

[9] S. Zhang, S. Zhang, T. Huang, and W. Gao, "Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching," *IEEE Transactions on Multimedia*, vol. 20, no. 6, pp. 1576–1590, June 2018.

[10] M. Tahon and L. Devillers, "Towards a small set of robust acoustic features for emotion recognition: Challenges," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 24, no. 1, pp. 16–28, 2016.

[11] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, and B. Weiss, "A database of german emotional speech," in *Proc. INTERSPEECH, 2005*, 2005, pp. 1517–1520.

[12] K. R. Scherer, "Vocal communication of emotion: A review of research paradigms," *Speech communication*, vol. 40, no. 1, pp. 227–256, 2003.

[13] G. Zhou, J. H. Hansen, and J. F. Kaiser, "Nonlinear feature based classification of speech under stress," *IEEE Transactions on Speech and Audio Processing*, vol. 9, no. 3, pp. 201–216, 2001.

[14] H. M. Teager, "Some observations on oral air flow during phonation," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 28, no. 5, pp. 599–601, 1980.

[15] N. Wang, P. Ching, N. Zheng, and T. Lee, "Robust speaker recognition using denoised vocal source and vocal tract features," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 1, pp. 196–205, 2011.

[16] R. Sant Ana, R. Coelho, and A. Alcaim, "Text-independent speaker recognition based on the hurst parameter and the multidimensional fractional brownian motion model," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 3, pp. 931–940, 2006.

[17] S. Wu, T. H. Falk, and W.-Y. Chan, "Automatic speech emotion recognition using modulation spectral features," *Speech Communication*, vol. 53, no. 5, pp. 768–785, 2011.

[18] K. Wang, N. An, B. N. Li, Y. Zhang, and L. Li, "Speech emotion recognition using fourier parameters," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 69–75, 2015.

[19] F. Eyben, K. R. Scherer, B. W. Schuller, J. Sundberg, E. André, C. Busso, L. Y. Devillers, J. Epps, P. Laukka, S. S. Narayanan, and K. Truong, "The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing," *IEEE Transactions on Affective Computing*, vol. 7, no. 2, pp. 190–202, 2016.

[20] N. Huang, Z. Shen, S. Long, M. Wu, H. Shih, Q. Zheng, N. Yen, C. Tung, and H. Liu, "The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis," *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 454, no. 1971, pp. 903–995, March 1998.

[21] H. E. Hurst, "Long-term storage capacity of reservoirs," *Trans. Amer. Soc. Civil Eng.*, vol. 116, pp. 770–808, 1951.

[22] P. Borgnat, P. Flandrin, P. Honeine, C. Richard, and J. Xiao, "Testing stationarity with surrogates: A time-frequency approach," *IEEE Transactions on Signal Processing*, vol. 58, no. 7, pp. 3459–3470, July 2010.

[23] D. Wu, J. Li, and H. Wu, "α-gaussian mixture modelling for speaker recognition," *Pattern Recognition Letters*, vol. 30, no. 6, pp. 589–594, 2009.

[24] D. A. Reynolds and R. C. Rose, "Robust text-independent speaker identification using gaussian mixture speaker models," *IEEE Transactions on Speech and Audio Processing*, vol. 3, no. 1, pp. 72–83, 1995.

[25] L. Rabiner and B. Juang, "An introduction to hidden markov models," *IEEE ASSP Magazine*, vol. 3, no. 1, pp. 4–16, 1986.

[26] C. Cortes and V. Vapnik, "Support vector networks," *Machine learning*, vol. 20, pp. 273–297, 1995.

[27] L. Deng and D. Yu, "Deep learning: Methods and applications," Tech. Rep., May 2014. [Online]. Available: https://www.microsoft.com/en-us/research/publication/deep-learning-methods-and-applications/

[28] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, November 1998.

[29] E. Çakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, June 2017.

[30] Z. Wu and N. Huang, "Ensemble empirical mode decomposition: a noise-assisted data analysis method," *Advances in Adaptive Data Analysis*, vol. 1, no. 1, pp. 1–41, 2009.

[31] A. Venturini, L. Zão, and R. Coelho, "On speech features fusion, α-integration gaussian modeling and multi-style training for noise robust speaker classification," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 22, no. 12, pp. 1951–1964, 2014.

[32] D. Veitch and P. Abry, "A wavelet-based joint estimator of the parameters of long-range dependence," *IEEE Transactions on Information Theory*, vol. 45, no. 3, pp. 878–897, 1999.

[33] I. Daubechies, *Ten lectures on wavelets*. Society for Industrial and Applied Mathematics, 1992, vol. 61.

[34] L. E. Baum, "An inequality and associated maximization thechnique in statistical estimation for probabilistic functions of markov process," *Inequalities*, vol. 3, pp. 1–8, 1972.

[35] A. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm," *IEEE Transactions on Information Theory*, vol. 13, no. 2, pp. 260–269, 1967.

[36] A. Milton, S. S. Roy, and S. T. Selvi, "SVM scheme for speech emotion recognition using MFCC feature," *International Journal of Computer Applications*, vol. 69, no. 9, 2013.

[37] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A.-r. Mohamed, N. Jaitly, A. Senior, V. Vanhoucke, P. Nguyen, T. N. Sainath, and B. Kingsbury, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[38] Y. Wang, A. Narayanan, and D. Wang, "On training targets for supervised speech separation," *IEEE/ACM Transactions on Audio, Speech and Language Processing*, vol. 22, no. 12, pp. 1849–1858, 2014.

[39] D. L. Wang and J. Chen, "Supervised speech separation based on deep learning: An overview," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 26, no. 10, pp. 1702–1726, 2018.

[40] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP)*, September 2015, pp. 1–6.

[41] J. Salamon and J. P. Bello, "Deep convolutional neural network and data augmentation for environmental sound classification," *IEEE Signal Processing Letters*, vol. 24, no. 3, pp. 279–283, March 2017.

[42] H. Zhang, I. McLoughlin, and Y. Song, "Robust sound event recognition using convolutional neural networks," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, April 2015, pp. 559–563.

[43] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2017, pp. 2392–2396.

[44] S. Young, G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey *et al.*, "The HTK book," *Cambridge university engineering department*, vol. 3, p. 175, 2002.

[45] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, p. 27, 2011.

[46] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, no. 4, pp. 335–359, 2008.

[47] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, "The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, no. 1, pp. 5–17, 2012.