IET Signal Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Hilbert–Huang–Hurst-based non-linear acoustic feature vector for emotion classification with stochastic models and learning systems

ISSN 1751-9675
Received on 20th August 2019
Revised 15th July 2020
Accepted on 22nd July 2020
E-First on 7th September 2020
doi: 10.1049/iet-spr.2019.0383
www.ietdl.org

Vinícius Vieira[1], Rosângela Coelho[2] ✉, Francisco Marcos de Assis[1]

[1]Post-Graduate Program in Electrical Engineering, Federal University of Campina Grande (UFCG), Campina Grande 58429-900, Brazil
[2]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*

Abstract: This study presents a widespread analysis of affective vocal expression classification systems. In this study, the Hilbert–Huang–Hurst coefficient (HHHC) vector is proposed as a non-linear vocal source feature to represent the emotional states according to their effects on the speech production mechanism. Affective states are highlighted by the empirical mode decomposition-based method, which exploits the non-stationarity of the acoustic variations. Hurst coefficients are then estimated from the decomposition modes to form the feature vector. Additionally, a vector of the index of non-stationarity (INS) is introduced as dynamic information to the HHHC. The proposed feature vector is evaluated in speech emotion classification experiments with three databases in German and English languages. Three state-of-the-art acoustic feature vectors are adopted as a baseline. The α-integrated Gaussian mixture model (α-GMM) is also introduced for the emotion representation and classification. Its performance is compared to competing for stochastic and machine learning classifiers. Results demonstrate that the HHHC leads to significant classification improvement when compared to the baseline acoustic feature vectors. Moreover, results also show that the α-GMM outperforms the competing classification methods. Finally, the complementarity aspects of HHHC and INS are also evaluated for the GeMAPS and eGeMAPS feature sets.

1 Introduction

Affective states play an important role in the cognition, perception, and communication of the human-being daily life. For instance, some unexpected events can motivate the occurrence of a happiness state, while stressful situations may cause health problems. Automatic emotion recognition is especially important to improve communication between humans and machines [1, 2]. In the literature, emotions are generally classified using physical or physiological signals such as speech [3], facial expression [4], and electrocardiogram (ECG) [5]. Particularly, speech emotion recognition has received much research attention in the past few years [6–11]. In this scenario, many promising applications can be considered, such as security access, automatic translation, call-centres, mobile communication, and human–robot interaction [12].

The speech production submitted to an emotional state is affected by changes in muscle tension and breathing rate. These changes lead to different speech signals depending on the emotion. Fig. 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) [13] 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 [14]. According to Ekman [2], there are certain emotions that can be naturally recognised by humans. Although there is a universality of the affective states' discrimination, their decoding in the computational field is difficult. The identification of an *affective vocal print* is fundamental to achieve a powerful

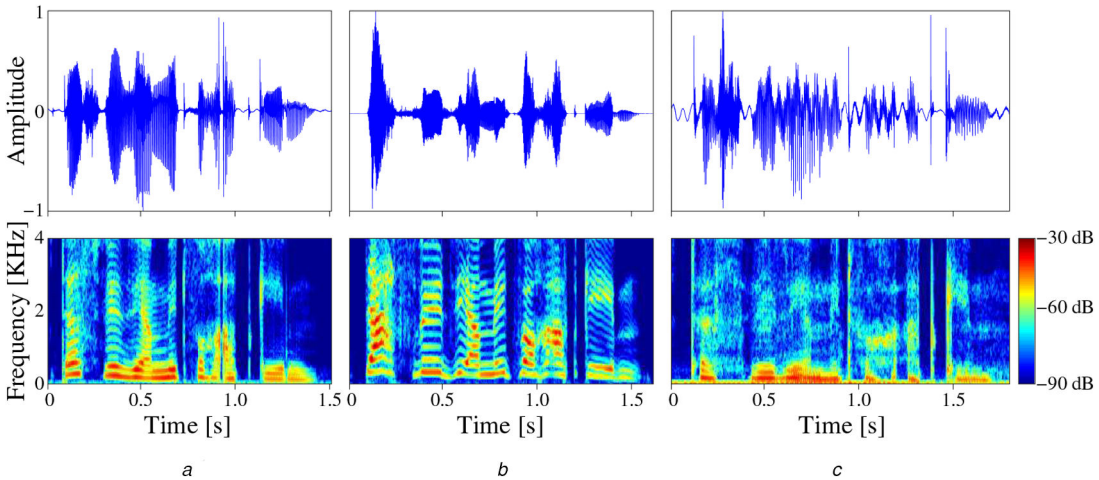

Fig. 1 *Amplitudes and spectrograms of speech signals produced considering different emotional states*
(a) Neutral, ***(b)*** Anger, ***(c)*** Sadness. These three signals correspond to the same female person speaking the same message in German: 'Das will sie am Mittwoch abgeben'

emotion recognition system. Thus, a key challenge is to define a feature that enables the characterisation of different emotions [3, 12]. 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 behaviour of multiple affective states is a crucial search.

In [15], Teager-Energy-Operator (TEO) [16] based features were proposed for the classification of stress conditions. The idea was to capture non-linear airflow structures of the acoustic signal induced by the speakers emotional state. Based on the fact that the excitation source signal reflects the speaker physiological behaviour, vocal source features may also be applied for this purpose. Such features are less dependent on the linguistic content of speech [17] in comparison to spectral ones. In [8], the pH vocal source feature vector [18] was evaluated for emotion and stress classification. The authors showed that TEO feature vectors might not be suitable for emotion classification. Both pH and TEO feature vectors do not take into account the non-linear effect of speech production, such as the non-stationarity of the affective acoustic variation and its dynamic behaviour. These aspects are important to be exploited by an acoustic affective attribute.

One of the most common feature vectors applied as a baseline in the literature and challenges is formed by mel-frequency cepstral coefficients (MFCC) [19]. This vector has been widely used for affective recognition [20] due to its success in other tasks, such as speech and speaker recognition [17, 21]. Nonetheless, other proposed features have shown superior performance than MFCC [8, 15, 21, 22]. For instance, pH [18] achieves 6.8 percentage points (p.p.) higher accuracy than MFCC in emotion classification [8]. Some approaches have focused on recognition rates improvement, where several features are combined to form collections of low-level descriptors (LLDs) [12, 23]. This means that there is not yet a consensus of an explicit single 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 individual representation of each affective state, which can improve the performance of classification tasks.

Stochastic classifiers such as Gaussian mixture model (GMM) [24] and hidden Markov model (HMM) [25] were widely adopted for speaker and speech recognition tasks. The use of such classifiers in emotion recognition [3, 7] is mainly due to their success in these speech applications. More recently, machine learning solutions have also been successfully adopted for speech emotion classification [10, 26]. Among others, these classification approaches include support vector machine (SVM) [27], deep neural network (DNN) [28], convolutional neural network (CNN) [29], convolutional recurrent neural network (CRNN) [30], and long short-term memory (LSTM) [31]. Until now, there is no consensus about which is the most suitable classifier for speech emotion recognition. Due to this fact, current research works still consider stochastic and machine learning approaches for emotion classification [20].

The main contribution of this work is the introduction of a new non-linear acoustic feature vector based on the non-stationary effects of emotions. The empirical mode decomposition (EMD) [32] is first applied to decompose the speech signals into a series of intrinsic mode functions (IMFs). Then, Hurst coefficients (H) [33] are estimated from each IMF on a frame-by-frame basis to compose the Hilbert–Huang–Hurst coefficient (HHHC) affective vector. In this proposal, the EMD is used to emphasise acoustic variations present in the speech signal, while Hurst coefficients can characterise highlighted vocal source components. It means that the combination of EMD with Hurst can capture the non-stationary acoustic variations that occur during speech production, which depend on the affective states. This aspect is still not well explored in the literature.

The index of non-stationarity (INS) [34] is here proposed as additional information to the HHHC feature vector. It dynamically describes the non-stationary behaviour of affective speech samples. The $\alpha-$ integrated GMM (α-GMM) [35] is also introduced to classify emotional states. It is compared to classic GMM and HMM stochastic methods, and also machine learning approaches SVM, DNN, CNN, and CRNN. Several experiments are conducted to show the effectiveness of the new vocal source feature vector in different languages and scenarios. EMO-DB [13], IEMOCAP (interactive emotional dyadic motion capture) [36], and SEMAINE (sustained emotionally coloured machine–human interaction using non-verbal expression) [37] databases are adopted for this purpose. Results demonstrate that the 6-dimensional HHHC vector is a pure and robust attribute for emotion. Additionally, HHHC vectors contribute as complementary to the Geneva Minimalistic Acoustic Parameter Set (GeMAPS) and its extended version (eGeMAPS) [23] to improve the classification rates.

This paper is organised as follows. Section 2 introduces the HHHC vector and presents the feature extraction procedure. The INS is also described in this section. The α-GMM and competing classifiers are presented in Section 3. Evaluation experiments are described in Section 4 and the results are exhibited in Section 5. Finally, Section 6 concludes this work.

2 New non-linear acoustic feature vector

The general idea of the HHHC vector is to characterise the vocal source when affected by an emotional state. The affective content of the speech is highlighted by the decomposition method of the Hilbert–Huang transform (HHT). Instead of the original EMD, the ensemble EMD (EEMD) [38] is applied to achieve improvement in the affective states detection. After the decomposition, Hurst coefficients, which are closely related to the excitation source, capture the non-linear information from the emphasised acoustic variations. In [39], it was shown that acoustic sources have different degrees of non-stationarity. In this work, a vector of INS values is proposed to analyse and detect speech emotional states.

2.1 HHHC feature vector

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

2.1.1 EMD/EEMD: EMD was introduced in [32] as a non-linear time-domain adaptive method for decomposing non-stationary signals into a series of oscillatory modes. As stated in [32], the EMD is the 'key part' of the HHT analysis, and it was proposed specifically for the HHT. The general idea is to locally analyse 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 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 adopted in this work can be summarised by 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 $\left(e_{\text{up}}(t)\right)$ and lower $(e_{\text{lo}}(t))$ envelopes, respectively.
(3) Define the local trend as

$$a(t) = \left(e_{\text{up}}(t) + e_{\text{lo}}(t)\right)/2. \quad (1)$$

(4) Calculate the detail component as $d(t) = x(t) - a(t)$.

Every IMF has zero mean, and the number 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 iteratively repeated with $x(t)$ replaced by $d(t)$ until the new detail can be considered as an IMF. In this work, this sifting process is repeated until the new IMF achieves the stopping criteria defined in [40]. 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 into a finite number of IMFs, the integrability property of the EMD can be expressed as

$$x(t) = \sum_{m=1}^{M} \mathrm{IMF}_m(t) + r(t), \tag{2}$$

where $r(t)$ is the last residual sequence.

As an alternative for the EMD, the EEMD method was proposed to avoid the *mode mixing* phenomena [38], which refer to IMF fluctuations that do not appear in the proper scale. Since these oscillations impact on the Hurst values estimated from the IMFs, the EEMD approach is expected to more properly emphasise affective acoustic variations than the EMD. Given the target signal $x(t)$, the EEMD method firstly generates an ensemble of I trials, $x^i(t)$, $i = 1, \ldots, I$, 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, $\mathrm{IMF}_m^i(t)$, $m = 1, \ldots, M$. Then, the mth mode of $x(t)$ is obtained as the average of the I corresponding IMFs.

Fig. 2 shows the EEMD applied to three speech segments of 40 ms collected from EMO-DB [13]. These segments refer to neutral speech (Fig. 2*a*) and two basic emotions: anger (Fig. 2*b*) and sadness (Fig. 2*c*). The EEMD applies a high-frequency versus low-frequency separation between IMFs. Note that the affective signals have different non-stationary dynamic behaviours. For instance, IMFs 1 and 2 of anger present amplitude values higher than the corresponding modes of the other signals. On the other hand, the highest amplitude values are observed in the last 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).

2.1.2 Hurst coefficients: The Hurst exponent $(0 < H < 1)$, or Hurst coefficient, expresses the time-dependence or scaling degree of a stochastic process [33]. Let a speech signal be represented by a stochastic process $x(t)$, with the normalised autocorrelation coefficient function $\rho(k)$, the H exponent is defined by the asymptotic behaviour of $\rho(k)$ as $k \to \infty$, i.e.

$$\rho(k) \sim H(2H-1)k^{2(H-2)}. \tag{3}$$

In this study, the H values are estimated from IMFs on a frame-by-frame basis using the wavelet-based estimator [41], 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 $(j = 1, 2, \ldots, J)$ 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 the detail coefficients, where N_j is the number of available coefficients for each scale j. In [41], 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 $y_j = \log_2(\sigma_j^2)$ 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 H values close to the unity. The authors extracted Hurst coefficients directly from the speech signal in a frame-basis for the pH feature vector [8]. In contrast, this present work deals with the estimation of Hurst values from the IMFs of speech signals.

The composition of HHHC vectors obtained from speech samples is illustrated in Fig. 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 [42] 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 analysed IMFs, high-arousal emotions have the lowest H averages.

2.1.3 HHHC feature extraction: The HHHC extraction from 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 three values of H is presented in Fig. 4. The decomposition is applied to each segment of the input signal. The Hurst coefficients are obtained on a frame-by-frame basis from each IMF. Then, the HHHC feature matrix is formed by concatenating the extracted acoustic feature vectors.

2.2 INS vector

The INS is a time–frequency approach to objectively examine the non-stationarity of a signal [34]. 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, a dissimilarity divergence $D(\cdot,\cdot)$ is used to obtain the distance between the spectrum of the analysed signal and its global spectrum averaged

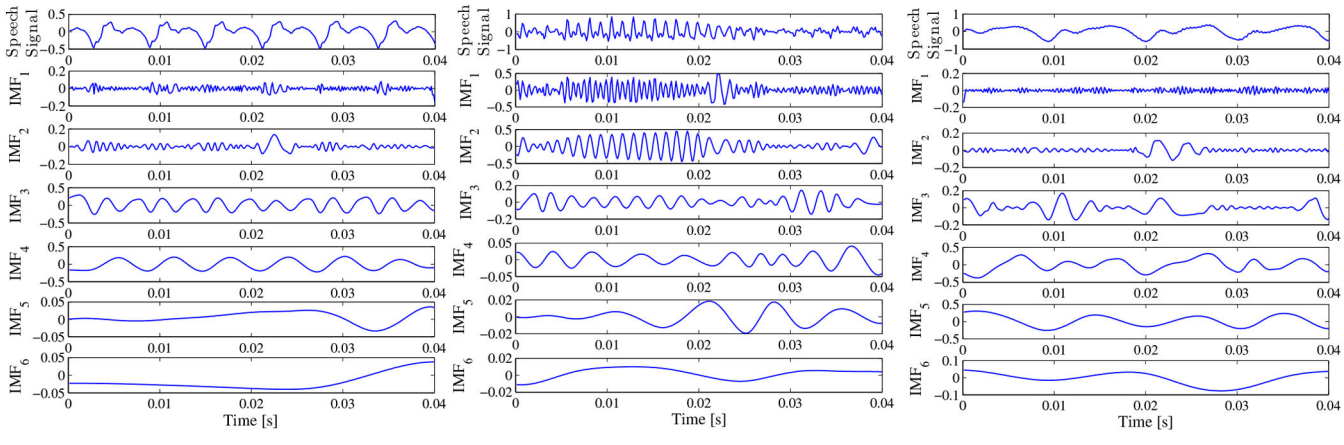

Fig. 2 *First six IMFs obtained with EEMD from voiced speech segments*
(a) Neutral, ***(b)*** Anger, ***(c)*** Sadness

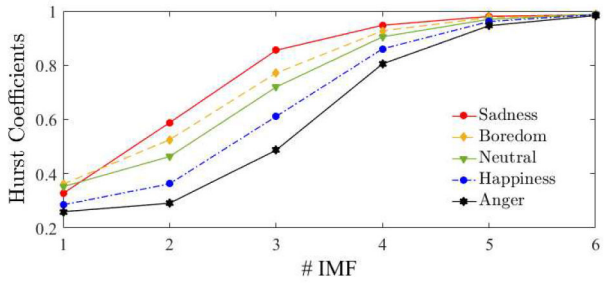

Fig. 3 *Hurst mean values of six IMFs obtained from speech samples under five non-stationary emotional variations*

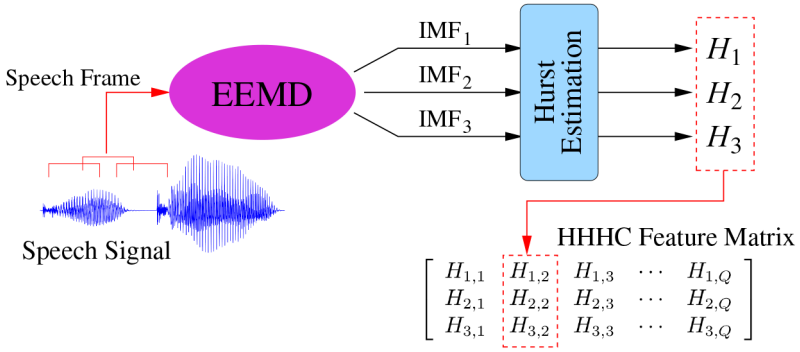

Fig. 4 *Example of an HHHC vector extraction with three coefficients*

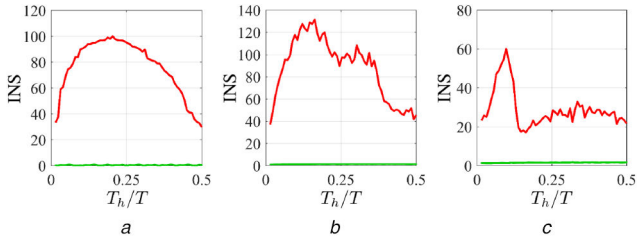

Fig. 5 *INS computed from voiced segments considering emotional states*
(a) Neutral, ***(b)*** Anger, ***(c)*** Sadness

over time. In [34], the authors propose the use of the divergence measure $D(p_1, p_2)$ between two distributions p_1 and p_2 defined as

$$D(p_1, p_2) = D_{\mathrm{KL}}(p_1, p_2)\,(1 + D_{\mathrm{LSD}}(p_1, p_2))\,, \tag{4}$$

where $D_{\mathrm{LSD}}(p_1, p_2)$ refers to the log-spectral deviation and $D_{\mathrm{KL}}(p_1, p_2)$ is the Kullback–Leibler divergence between normalised versions of p_1 and p_2. Let $D_n^{(x)}$ denote the divergence of the spectrogram of the analysed signal computed at the time position $t_n (n = 1, \ldots, N)$. Similarly, $D_n^{(sj)}$ denotes the distance measured from the j surrogate sequence $(n = 1, \ldots, N, \quad j = 1, \ldots, J)$. Then, variances are obtained from the divergence values as

$$\begin{cases} \Theta_0(j) = \mathrm{var}\left(D_n^{(sj)}\right)_{n=1,\ldots,N}, \; j = 1, \ldots, J \\ \Theta_1 = \mathrm{var}\left(D_n^{(x)}\right)_{n=1,\ldots,N} \end{cases}. \tag{5}$$

Finally, the INS is given by

$$\mathrm{INS} := \sqrt{\Theta_1 / \langle \Theta_0(j) \rangle_j}, \tag{6}$$

where $\langle \cdot \rangle$ is the mean value of $\Theta_0(j)$. In [34], the authors considered that the distribution of the divergence values could 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

$$\mathrm{INS} \begin{cases} \leq \gamma, & \text{signal is stationary}, \\ > \gamma, & \text{signal is non-stationary}\,. \end{cases} \tag{7}$$

Fig. 5 depicts examples of the INS obtained from voiced segments selected from the EMO-DB. Once again, these segments correspond to the neutral state and two emotional variations: anger and sadness. The time scale $T_h/T \in [0.0015, 0.5]$ 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. For each signal, a total of $J = 50$ surrogates are randomly generated for the INS computation. 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 most of the observed time scales, the INS of Sadness reaches a maximum value of 60. On the other hand, Anger presents INS greater than 100 for several time scales.

Figs. 3 and 5 indicate that, although HHHC and INS are based on different time–frequency analysis methods, both can capture relevant information regarding the speaker emotional state. In this work, the INS is computed from each decomposition mode to capture the non-stationarity dynamics of each IMF. Thus, INS values are expected to reflect complementary dynamic information to the HHHC vector. Due to this reason, the INS vector is here proposed to be used together with the HHHC acoustic feature vector, which is hereinafter denoted as HHHC + INS.

The procedure adopted in this work to obtain the INS vectors can be summarised in the following steps:

(1) Apply the EEMD to decompose the target speech signal into a series of M modes $\mathrm{IMF}_m(t), m = 1, \ldots, M$.
(2) Given a set of D time scale values, compute the INS from each mode $\mathrm{IMF}_m(t)$. Thus, a vector of D INS values, $\mathbf{INS}_m$, is composed of each $m = 1, \ldots, M$.
(3) Vectors $\mathbf{INS}_m, m = 1, \ldots, M$ are concatenated to form a single INS vector with DM coefficients.

3 Classification task

The α-GMM is here proposed for acoustic emotion classification. The α-GMM was firstly adopted for speaker identification [35]. By introducing a factor of α, the modelling capacity of the GMM is extended, which is more suitable for acoustic variations conditions.

The α-integration generalises the linear combination of the conventional GMM ($\alpha = -1$). For $\alpha < -1$, the α-GMM classifier emphasises larger probability values and de-emphasises 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. In accordance with [35], it was demonstrated in [39] that α-GMM outperforms the conventional GMM. Hence, the HHHC vector is evaluated considering the α-GMM and the classical GMM ($\alpha = -1$). Five other classifiers are used for comparative evaluation purposes.

3.1 α-integrated Gaussian mixture model

Given an affective state model λ_L, composed of M Gaussian densities $b_i(\boldsymbol{x})$, $i = 1, \ldots, M$, the α-integration of densities is defined as [35]

$$p(\boldsymbol{x}|\lambda_L) = C\left[\sum_{i=1}^{M} \pi_i b_i(\boldsymbol{x})^{\frac{1-\alpha}{2}}\right]^{\frac{2}{1-\alpha}}, \tag{8}$$

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

Models λ_L are completely parameterised by mean vectors, covariance matrices, and weights of Gaussian densities. These parameters are estimated using an adapted expectation-maximisation (EM) algorithm to maximise the likelihood function

$$p(\boldsymbol{X}|\lambda_L) = \prod_{t=1}^{Q} p(\boldsymbol{x}_t|\lambda_L), \tag{9}$$

where $\boldsymbol{X} = [\boldsymbol{x}_1\boldsymbol{x}_2 \ldots \boldsymbol{x}_Q]$ is the feature matrix extracted from the training speech segment Φ_L of the affective state L. During tests, the speaker emotion is identified according to the maximum likelihood criterion. It means that the classified emotion L corresponds to the model λ_L that maximises the likelihood function in (9).

3.2 Hidden Markov models

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

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

The standard algorithm for HMM training is the forward-backward, or Baum–Welch algorithm [43]. It obtains matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ that maximise the likelihood $p(\boldsymbol{x}|\lambda_L)$. The Viterbi algorithm is commonly used for decoding [44].

3.3 Support vector machines

SVM [27] is a classical supervised machine learning model widely applied for data classification. The general idea is to find the optimal separating hyperplane, which maximises the margin on the training data. For this purpose, it transforms input vectors into a high-dimensional feature space using a non-linear transformation (with a kernel function) space. Given a training set $\{u_\xi\}_{\xi=1}^{N} = \{(\boldsymbol{x}_\xi, L_\xi)\}_{\xi=1}^{N}$, where $L_\xi \in \{-1, +1\}$ represents the affective state L of the utterance ξ. Thus, the classifier is a hyperplane defined as $g(\boldsymbol{x}) = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b$, 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 to the utterance, can be indicated by $L_\xi g(\boldsymbol{x}_\xi)$. For $L_\xi = +1$, $L_\xi g(\boldsymbol{x}_\xi)$ must be greater than 1, while $L_\xi g(\boldsymbol{x}_\xi)$ is required to be smaller than -1 for $L_\xi = -1$. Then, the hyperplane is chosen by the solution of the optimisation problem of minimising $\frac{1}{2}\boldsymbol{w}^{\mathrm{T}}\boldsymbol{w}$ subject to $L_\xi(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x} + b) \geq 1, \xi = 1, 2, \ldots, N$.

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 [26]. Radial basis function (RBF) is used as the SVM kernel.

3.4 Deep neural networks

DNN is one of the most prominent methods for machine learning tasks such as speech recognition [45], separation [46], and emotion classification [10]. The deep learning concept can be applied for architectures such as feedfoward multilayer perceptrons (MLPs), CNNs and RNNs [47]. 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 [45].

The majority voting strategy is adopted for the emotion classification with DNN. It means that the DNN is first applied to classify each frame vector to an emotional condition. Then, the emotion assigned to the entire speech segment is the one that received the maximum number of frame labels.

3.5 Convolutional neural networks

CNNs [29] have been widely adopted in the acoustic signal processing area, particularly for sound classification [48, 49] and sound event detection [50]. CNNs extend the multilayer perceptrons model by introducing 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.

3.6 Convolutional recurrent neural networks

CRNNs [30] consist of the combination of CNNs with RNN. The idea is to improve the CNN by learning spectro-temporal information of relatively longer events that are not captured by convolutional layers. For this purpose, recurrent layers are applied to the output of the convolutional layer to integrate the information of earlier time windows. In the literature, CNNs and RNNs have been successfully combined for music classification [51] and sound event detection [30]. In this work, a single feedforward layer with a sigmoid activation function that follows the recurrent layers is considered as the output layer of the network [30].

4 Experimental setup

Extensive experiments are carried out to evaluate the proposed HHHC acoustic feature vector. In the training phase, affective models are generated after the pre-processing and feature extraction steps. 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. For the tests, 800 ms speech segments are applied for each emotion of the testing speaker. The detection of emotional content in instances with <1 s time duration is suitable for real-life situations [12].

The α-GMM is evaluated 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 [52] with the left-to-right topology. For each affective condition, five HMM states are used

with one single Gaussian each. The SVM implementation is carried out with the LIBSVM [53], 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 setup adopted in this paper is according to the DNN configuration presented in [46], considering multilayer perceptrons with three hidden layers. The networks are trained with the standard backpropagation algorithm with dropout regularisation (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. The hidden layers are composed of 1024 rectified linear units each. 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 [30]. A single recurrent layer is used to compose the CRNN.

In order to verify the classification rates improvement for emotion recognition, the proposed HHHC vector is also evaluated as complementary to collections of features such as GeMAPS [23]. For this purpose, binary arousal and valence classification experiments are carried out using the SVM classifier.

4.1 Speech emotion databases

Three databases are considered in the experiments: EMO-DB [13], IEMOCAP (Interactive Emotional Dyadic Motion Capture) [36], and SEMAINE (Sustained Emotionally coloured Machine-human Interaction using Nonverbal Expression) [37]. Only the voiced segments of speech are considered in the experiments. For this purpose, the pre-processing step selects frames of 16 ms with high energy and a low zero-crossing rate. The sampling rate used for all databases is 8 kHz.

EMO-DB consists of ten actors (five women and five 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 vector in characterise emotions that are naturally recognised by humans. Thus, five emotions were chosen to show the effectiveness of the HHHC vector. A total of 40 s of voiced speech segments is available for each emotional state.

IEMOCAP is composed of both scripted and spontaneous conversations in the English language. Ten actors (five women and five men) were recorded in dyadic sessions in order to provide a more natural interaction of the targeted emotion. Although it comprises 12 h of recordings, a portion of the IEMOCAP database is applied to obtain the short emotional instances to be used for tests. Four emotional states are considered: anger, happiness, neutral, and sadness. For each emotional state, a total of 10 min of voiced content are used in the experiments: 5 min of scripted and 5 min of spontaneous speech.

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). Recordings of ten participants (five women and five men) are chosen for the experiments. From 27 categories (styles), four emotional states are selected: anger, happiness, amusement, and sadness. The set of voiced speech samples for each emotional state has 90 s.

4.2 Extracted features

Six-dimensional HHHC vectors are extracted according to the procedure presented in Section 2.1. For the EEMD-based analysis, the number of ensembles is set to $I = 100$. A total of 11 Gaussian noise levels are evaluated 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 vectors are used in the experiments. Fusion procedures are carried out for an improvement provided by the proposed HHHC in the recognition rates of the baseline features vectors.

4.2.1 MFCC: The extraction of mel-frequency cepstral coefficients [19] starts with the computation of the fast Fourier transform (FFT) from short-time frames of the speech signal. Mel-scaled bandpass filters are then used to obtain the spectral envelope of each frame. The mel scale better represents the human auditory system when compared to the linear scale. Frequencies in mel (f_{Mel}) and linear (f_{Hz}) scales are related by

$$f_{\text{Mel}} = 1127 \log\left(1 + \frac{f_{\text{Hz}}}{700}\right). \tag{10}$$

The discrete cosine transform is finally applied to the output of the mel scaled filters to achieve the MFCC vectors. In this work, 12-dimensional MFCC vectors are obtained from speech frames of 25 ms, with a frame rate of 10 ms.

4.2.2 TEO-CB-Auto-Env: The TEO [54] was developed to reflect the non-linear energy flow within the vocal tract during the speech production. The idea is to capture the vortex-flow interactions induced by changes in the vocal system due to emotional states. For a discrete-time signal $x(t)$, the TEO (Ψ) is given by

$$\Psi(x(t)) = x^2(t) - x(t+1)x(t-1). \tag{11}$$

The extractor of the critical band based TEO autocorrelation envelope (TEO-CB-Auto-Env) first splits the speech signal into critical bands using Gabor bandpass filters. The TEO is applied to capture the non-linear energy flow of each frequency band. For each frame, normalised autocorrelation functions are computed and the corresponding areas under the curves are used to compose the TEO-based feature vector. In this work, TEO-CB-Auto-Env vectors with 16 coefficients are extracted from 75 ms speech samples, with 50% overlapping.

4.2.3 pH: The pH vocal source feature vector was proposed in [18] for automatic speaker classification. It is composed of Hurst values that express the scaling degree of the analysed signal. In [8], the authors showed that H values of speech signals produced in high-arousal emotions generally rely on the range $0 < H < 1/2$, while low-arousal emotions lead to $1/2 < H < 1$. Moreover, pH outperformed MFCC and TEO-CB-Auto-Env feature vectors in speech emotion recognition experiments.

The pH extractor applies the DWT to successively decompose the speech signal into sequences of approximation and detail coefficients. The wavelet-based estimator is then applied to obtain H values from the approximation sequences. For each time frame, a pH vector is composed of Hurst values estimated from the original samples and all decomposition scales. For the experiments, the estimation of the pH feature vector is conducted considering frames of 50 ms, every 10 ms, using the Daubechies wavelet filters with 12 coefficients (2–12 scales).

5 Results

This section presents accuracies results obtained in speech emotion classification. Confusion matrices achieved for the EMO-DB, IEMOCAP, and SEMAINE databases are shown in Tables 1–3, respectively. These confusion matrices are attained with α-GMM, HMM, and SVM classifiers for the HHHC, HHHC + INS, and baseline feature vectors. Considering the Chebyshev inequality [55] and a confidence degree of 95%, the precision obtained for the accuracy values due to the number of tests is 0.0070 for EMO-DB (250 tests), 0.0021 for IEMOCAP (2840 tests), and 0.0053 for SEMAINE (450 tests). Although the HHHC vector extracted with the conventional EMD outperforms the competing attributes, the EEMD-based approach reaches even higher accuracies. Results for

HHHC are achieved with the EEMD-based approach considering low Gaussian noise level ($0.005 \leq \text{std} \leq 0.02$).

5.1 Results with EMO-DB

For the α-GMM, the proposed HHHC vector achieves the highest average accuracy (79.2%) with three values of α (-4, -6 and -8). This result is better than that attained with pH for $\alpha = -2$ (65.4%). HHHC also outperform the MFCC (63.6%) and TEO (52.8%) feature vectors in 15.6. and 26.4 p.p., respectively. The INS information contributes to>2 p.p. for the HHHC average accuracy. The HHHC vector achieves 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.

Fig. 6 presents the average classification accuracies obtained with the proposed and baseline feature vectors considering the neural network classifiers. Average results obtained with the α-GMM are also shown in Fig. 6. Note that HHHC and HHHC + INS achieve the best results for all classifiers. For the CRNN, which outperforms DNN and CNN, the HHHC vector 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 fusion, e.g. the average accuracy with α-GMM is 2.6 p.p. greater than CRNN.

Fig. 7 shows the identification accuracy with α-GMM for the feature fusion between HHHC and competing for feature vectors. 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 the HHHC vector increases in almost 10 p.p. the recognition rate provided by the MFCC feature vector. Concerning the TEO + HHHC fusion, the best average accuracy is 72.1% with $\alpha = -6$ and $\alpha = -8$, which means an improvement of 19.2 p.p. for the TEO-based feature vector.

5.2 Results with IEMOCAP

It can be seen from Table 2 that, for all considered feature sets, the α-GMM leads to superior accuracies when compared to the HMM and the SVM classifiers. Only HHHC and HHHC + INS reach average accuracies higher than 60.0%. These values are achieved using the α-GMM with $\alpha = -8$. In comparison to baseline feature

Table 1 Accuracy rates (%) of five emotional states with the HHHC and baseline feature vectors for EMO-DB

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

Bold values correspond to the percentage of correctly classified emotions

vectors, the HHHC vector leads to an average accuracy 8 p.p. higher than pH vector ($\alpha = -8$), 10 p.p. higher than MFCC ($\alpha = -4$) and 15 p.p. higher than the TEO-based feature vector ($\alpha = -6$). For each emotional state, the α-GMM attains >50.0% accuracies with HHHC. Furthermore, the α-GMM provides an improved performance with the baseline feature vectors, in comparison to HMM and SVM approaches.

Fig. 8 presents the average classification accuracies of IEMOCAP considering the α-GMM and neural network classifiers. Similarly to the EMO-DB, the HHHC vector outperforms the pH, MFCC, and TEO feature vectors for all classifiers. For the CRNN, the HHHC vector achieves an average accuracy of 54.3%, which is 3.0, 7.0, and 12.0 p.p. greater than pH, MFCC, and TEO, respectively. Moreover, HHHC + INS leads to the best results in all scenarios. The α-GMM also outperforms the competing classifiers for all features sets.

Fig. 9 depicts the results achieved with the features fusion using the α-GMM for the HHHC and baseline feature vectors in the IEMOCAP database. The pH + HHHC fusion obtains an accuracy of 63.2% ($\alpha = -8$), which outperforms both pH (52.8%), and HHHC + INS (62.8%). The fusion of Hurst-based feature vectors (pH + HHHC) indicates that the relation between H and the excitation source enables high performance in the discrimination of basic emotions. In comparison to the MFCC, the MFCC + HHHC fusion improves the average accuracy from 50.8 to 60.5% ($\alpha = -4$). Considering the TEO + HHHC fusion, the best result (56.1%) is achieved with $\alpha = -4$, which is 11.9 p.p. higher than that obtained with the TEO-based feature vector only.

5.3 Results with SEMAINE

The best average accuracies are obtained with HHHC and HHHC + INS (refer to Table 3): 54.5 and 57.0%, respectively, using α-GMM with $\alpha = -6$. These results are greater than those obtained with the baseline feature vectors: 50.8% for pH ($\alpha = -4$), 49.0% for the MFCC ($\alpha = -6$), and 40.8% for the TEO-based feature vector ($\alpha = -8$). An important issue about the SEMAINE database is mainly concerned with the Happiness and Amusement states recognition. Although these emotions present similar behaviour, the HHHC shows to be able to recognise both of them with >50.0% classification accuracy with the α-GMM. For baseline feature vectors, the α-GMM average result reaches >4 p.p. over

Table 2 Accuracy rates (%) of four emotional states with the HHHC and baseline feature vectors for IEMOCAP

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

Bold values correspond to the percentage of correctly classified emotions

Table 3 Accuracy rates (%) of four emotional states with the HHHC and baseline feature vectors for SEMAINE

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

Bold values correspond to the percentage of correctly classified emotions

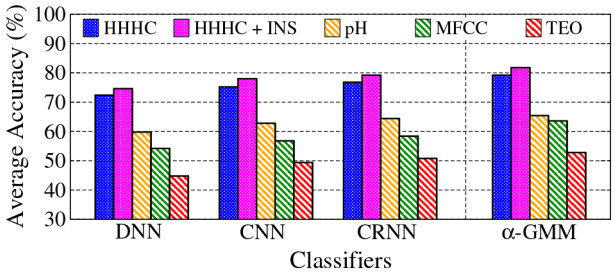

Fig. 6 *Average accuracies of EMO-DB obtained with* $\boldsymbol{\alpha}$*-GMM and neural network classifiers*

HMM and 10 p.p. over SVM. The α-GMM outperforms the HMM and SVM classifiers for all considered emotional states. According to the average classification results shown in Fig. 10, α-GMM also leads to the highest classification rates when compared to 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 feature vectors using α-GMM are shown in Fig. 11. The pH + HHHC fusion attains an average accuracy of 56.5%, which represents an improvement over the pH and HHHC feature vectors. With the MFCC + HHHC features fusion and $\alpha = -6$, the recognition rate varies from 49.0 to 53.6%. The HHHC provides an improvement of more than 6 p.p. when compared to the TEO-based feature vector (47.4%, $\alpha = -8$). The proposed feature vector is also very promising for discriminant learning strategies [10] applied to DNN and deep CNN methods for speech emotion classification.

5.4 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. According to the psychological dimensional theory presented in [56], arousal and valence are independent dimensions of affective states. The term arousal refers to the physiological activation level according to a person's emotional condition, varying from calm (or low) to excited (or high). In this work, anger, fear, and happiness are considered as high arousal emotions, while boredom, sadness, neutral, and disgust are low arousal

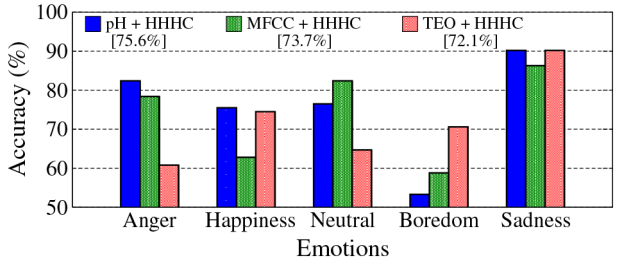

Fig. 7 *Classification accuracies with feature fusion and* $\boldsymbol{\alpha}$*-GMM classifier of emotional states from EMO-DB*

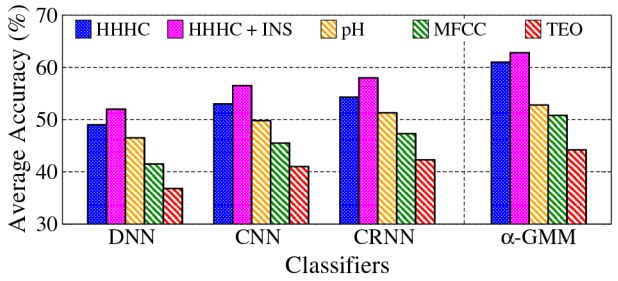

Fig. 8 *Average accuracies of IEMOCAP obtained with* $\boldsymbol{\alpha}$*-GMM and neural network classifiers*

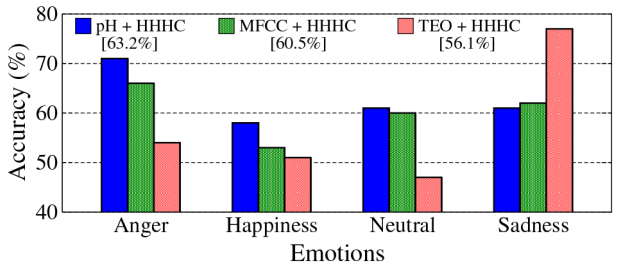

Fig. 9 *Classification accuracies with feature fusion and* $\boldsymbol{\alpha}$*-GMM classifier of emotional states from IEMOCAP*

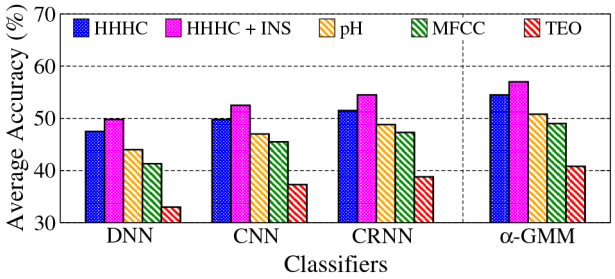

Fig. 10 *Average accuracies of SEMAINE obtained with* $\boldsymbol{\alpha}$*-GMM and neural network classifiers*

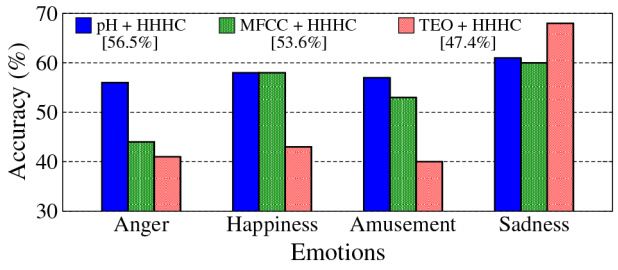

Fig. 11 *Classification accuracies with feature fusion and* $\boldsymbol{\alpha}$*-GMM classifier of emotional states from SEMAINE*

affective states. On the other hand, valence is related to the pleasantness induced by an affective state: pleasant (positive) or unpleasant (negative). In terms of valence, Happiness and Neutral are assumed as positive conditions while the remaining states are negative.

Table 4 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**

Bold values refer to the best results achieved with the GeMAPS and eGeMAPS features sets.

The GeMAPS feature set and its extended version (eGeMAPS) [23] are adopted for the binary classification experiments. The GeMAPS is formed of 62 functionals extracted from 18 low-level descriptors and six temporal features. Functionals of the other seven LLD are added to the GeMAPS feature set to compose the eGeMAPS, leading to a total of 88 parameters. The experimental setup is similar to [23] with eight folds cross-validation, 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 4.Table 4 shows the results of UAR (unweighted average recall) obtained from experiments with GeMAPS, eGeMAPS, HHHC, HHHC + INS, and the feature fusion of the proposed acoustic feature vector with the comparative feature sets. Note that, for arousal evaluation, GeMAPS and eGeMAPS reach >93% UAR while HHHC and HHHC + INS achieve 80.5 and 83.2%, respectively. While the standard feature sets need 62 and 88 parameters (GeMAPS and eGeMAPS, respectively) for this result, the HHHC vector shows interesting accuracy for a low-dimensional feature vector. However, HHHC and HHHC + INS contribute to an improvement in the UAR obtained with GeMAPS and eGeMAPS. For instance, the eGeMAPS + HHHC + INS fusion 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. It means that the HHHC and the HHHC + INS vectors can significantly improve the performance of large feature sets by adding only a few more parameters.

6 Conclusion

This work introduced the HHHC non-linear 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 feature vectors, HHHC obtained superior accuracy considering three different databases. On the feature fusion, HHHC vectors provide an improved performance for all considered baseline feature vectors. 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 vector showed that the HHHC vector is very promising for affective state representation and for classification tasks. Also, the HHHC complementarity to the GeMAPS features set was verified by the improvement in the recognition rates in binary arousal and valence emotion classification.

7 Acknowledgment

This work was supported in part by the National Council for Scientific and Technological Development (CNPq) under research grants nos. 140816/2014-3 and 307866/2015-7, and Fundação de Amparo à Pesquisa do Estado do Rio de Janeiro (FAPERJ) under the research grant no. 203075/2016.

8 References

[1] Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., *et al.*: 'Emotion recognition in human–computer interaction', *IEEE Signal Process. Mag.*, 2001, **18**, (1), pp. 32–80

[2] Ekman, P.: 'Basic emotions', in Dalgleish, T., Power, M., (Eds.): '*The handbook of cognition and emotion*' (Wiley Online Library, England, 1999), pp. 45–60

[3] El Ayadi, M., Kamel, M.S., Karray, F.: 'Survey on speech emotion recognition: features, classification schemes, and databases', *Pattern Recognit.*, 2011, **44**, (3), pp. 572–587

[4] Kalsum, T., Anwar, S.M., Majid, M., *et al.*: 'Emotion recognition from facial expressions using hybrid feature descriptors', *IET Image Process.*, 2018, **12**, (6), pp. 1004–1012

[5] Agrafioti, F., Hatzinakos, D., Anderson, A.K.: 'ECG pattern analysis for emotion detection', *IEEE Trans. Affective Comput.*, 2012, **3**, (1), pp. 102–115

[6] Tawari, A., Trivedi, M.M.: 'Speech emotion analysis: exploring the role of context', *IEEE Trans. Multimed.*, 2010, **12**, (6), pp. 502–509

[7] Schuller, B., Vlasenko, B., Eyben, F., *et al.*: 'Acoustic emotion recognition: a benchmark comparison of performances'. Proc. of the IEEE Workshop on Automatic Speech Recognition & Understanding, Merano, Italy, 2009, pp. 552–557

[8] Zão, L., Cavalcante, D., Coelho, R.: 'Time–frequency feature and AMS-GMM mask for acoustic emotion classification', *IEEE Signal Process. Lett.*, 2014, **21**, (5), pp. 620–624

[9] Huang, Y., Wu, A., Zhang, G., *et al.*: 'Extraction of adaptive wavelet packet filter-bank-based acoustic feature for speech emotion recognition', *IET Signal Process.*, 2015, **9**, (4), pp. 341–348

[10] Zhang, S., Zhang, S., Huang, T., *et al.*: 'Speech emotion recognition using deep convolutional neural network and discriminant temporal pyramid matching', *IEEE Trans. Multimed.*, 2018, **20**, (6), pp. 1576–1590

[11] Zhang, W., Song, P.: 'Transfer sparse discriminant subspace learning for cross-corpus speech emotion recognition', *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2020, **28**, pp. 307–318

[12] Tahon, M., Devillers, L.: 'Towards a small set of robust acoustic features for emotion recognition: challenges', *IEEE/ACM Trans. Audio, Speech Lang. Process.*, 2016, **24**, (1), pp. 16–28

[13] Burkhardt, F., Paeschke, A., Rolfes, M., *et al.*: 'A database of German emotional speech'. Proc. of the INTERSPEECH, Lisbon, Portugal, September 2005, pp. 1517–1520

[14] Scherer, K.R.: 'Vocal communication of emotion: a review of research paradigms', *Speech Commun.*, 2003, **40**, (1), pp. 227–256

[15] Zhou, G., Hansen, J.H., Kaiser, J.F.: 'Non-linear feature based classification of speech under stress', *IEEE Trans. Speech Audio Process.*, 2001, **9**, (3), pp. 201–216

[16] Teager, H.M.: 'Some observations on oral air flow during phonation', *IEEE Trans. Acoust. Speech Signal Process.*, 1980, **28**, (5), pp. 599–601

[17] Wang, N., Ching, P., Zheng, N., *et al.*: 'Robust speaker recognition using denoised vocal source and vocal tract features', *IEEE Trans. Audio, Speech, Lang. Process.*, 2011, **19**, (1), pp. 196–205

[18] Sant Ana, R., Coelho, R., Alcaim, A.: 'Text-independent speaker recognition based on the Hurst parameter and the multidimensional fractional Brownian motion model', *IEEE Trans. Audio, Speech, Lang. Process.*, 2006, **14**, (3), pp. 931–940

[19] Davis, S., Mermelstein, P.: 'Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences', *IEEE Trans. Acoust. Speech Signal Process.*, 1980, **28**, (4), pp. 357–366

[20] Mao, S., Tao, D., Zhang, G., *et al.*: 'Revisiting hidden Markov models for speech emotion recognition'. Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Brighton, UK, May 2019, pp. 6715–6719

[21] Wu, S., Falk, T.H., Chan, W.Y.: 'Automatic speech emotion recognition using modulation spectral features', *Speech Commun.*, 2011, **53**, (5), pp. 768–785

[22] Wang, K., An, N., Li, B.N., *et al.*: 'Speech emotion recognition using Fourier parameters', *IEEE Trans. Affective Comput.*, 2015, **6**, (1), pp. 69–75

[23] Eyben, F., Scherer, K.R., Schuller, B.W., *et al.*: 'The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing', *IEEE Trans. Affective Comput.*, 2016, **7**, (2), pp. 190–202

[24] Reynolds, D.A., Rose, R.C.: 'Robust text-independent speaker identification using Gaussian mixture speaker models', *IEEE Trans. Speech Audio Process.*, 1995, **3**, (1), pp. 72–83

[25] Rabiner, L., Juang, B.: 'An introduction to hidden Markov models', *IEEE ASSP Mag.*, 1986, **3**, (1), pp. 4–16

[26] Milton, A., Roy, S.S., Selvi, S.T.: 'SVM scheme for speech emotion recognition using MFCC feature', *Int. J. Comput. Appl.*, 2013, **69**, (9), pp. 34–39

[27] Cortes, C., Vapnik, V.: 'Support vector networks', *Mach. Learn.*, 1995, **20**, pp. 273–297

[28] Deng, L., Yu, D.: '*Deep learning: methods and applications*' (NOW Publishers, USA., 2014)

[29] Lecun, Y., Bottou, L., Bengio, Y., *et al.*: 'Gradient-based learning applied to document recognition', *Proc. IEEE*, 1998, **86**, (11), pp. 2278–2324

[30] Çakır, E., Parascandolo, G., Heittola, T., *et al.*: 'Convolutional recurrent neural networks for polyphonic sound event detection', *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2017, **25**, (6), pp. 1291–1303

[31] Hochreiter, S., Schmidhuber, J.: 'Long short-term memory', *Neural Comput.*, 1997, **9**, (8), pp. 1735–1780

[32] Huang, N., Shen, Z., Long, S., *et al.*: 'The empirical mode decomposition and the Hilbert spectrum for non-linear and non-stationary time series analysis', *Proc. R. Soc. Lond. A: Math., Phys. Eng. Sci.*, 1998, **454**, (1971), pp. 903–995

[33] Hurst, H.E.: 'Long-term storage capacity of reservoirs', *Trans. Am. Soc. Civil Eng.*, 1951, **116**, pp. 770–808

[34] Borgnat, P., Flandrin, P., Honeine, P., *et al.*: 'Testing stationarity with surrogates: a time–frequency approach', *IEEE Trans. Signal Process.*, 2010, **58**, (7), pp. 3459–3470

[35] Wu, D., Li, J., Wu, H.: 'α-Gaussian mixture modelling for speaker recognition', *Pattern Recognit. Lett.*, 2009, **30**, (6), pp. 589–594

[36] Busso, C., Bulut, M., Lee, C.C., *et al.*: 'IEMOCAP: interactive emotional dyadic motion capture database', *Lang. Res. Eval.*, 2008, **42**, (4), pp. 335–359

[37] McKeown, G., Valstar, M., Cowie, R., *et al.*: 'The SEMAINE database: annotated multimodal records of emotionally colored conversations between a person and a limited agent', *IEEE Trans. Affective Comput.*, 2012, **3**, (1), pp. 5–17

[38] Wu, Z., Huang, N.: 'Ensemble empirical mode decomposition: a noise-assisted data analysis method', *Adv. Adapt. Data. Anal.*, 2009, **1**, (1), pp. 1–41

[39] Venturini, A., Zão, L., Coelho, R.: 'On speech features fusion, α-integration Gaussian modeling and multi-style training for noise robust speaker classification', *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2014, **22**, (12), pp. 1951–1964

[40] Rilling, G., Flandrin, P., Goncalvès, P.: 'On empirical mode decomposition and its algorithms'. Proc. of the IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, Grado, Italy, 2003, pp. 444–447

[41] Veitch, D., Abry, P.: 'A wavelet-based joint estimator of the parameters of long-range dependence', *IEEE Trans. Inf. Theory*, 1999, **45**, (3), pp. 878–897

[42] Daubechies, I.: '*Ten lectures on wavelets*' (Society for Industrial and Applied Mathematics, USA., 1992), vol. **61**

[43] Baum, L.E.: 'An inequality and associated maximization technique in statistical estimation for probabilistic functions of Markov process', *Inequalities*, 1972, **3**, pp. 1–8

[44] Viterbi, A.: 'Error bounds for convolutional codes and an asymptotically optimum decoding algorithm', *IEEE Trans. Inf. Theory*, 1967, **13**, (2), pp. 260–269

[45] Hinton, G., Deng, L., Yu, D., *et al.*: 'Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups', *IEEE Signal Process. Mag.*, 2012, **29**, (6), pp. 82–97

[46] Wang, Y., Narayanan, A., Wang, D.: 'On training targets for supervised speech separation', *IEEE/ACM Trans. Audio, Speech Lang. Process.*, 2014, **22**, (12), pp. 1849–1858

[47] Wang, D.L., Chen, J.: 'Supervised speech separation based on deep learning: an overview', *IEEE/ACM Trans. Audio, Speech, Lang. Process.*, 2018, **26**, (10), pp. 1702–1726

[48] Piczak, K.J.: 'Environmental sound classification with convolutional neural networks'. Proc. of the IEEE 25th Int. Workshop on Machine Learning for Signal Processing, Boston, USA, September 2015, pp. 1–6

[49] Salamon, J., Bello, J.P.: 'Deep convolutional neural networks and data augmentation for environmental sound classification', *IEEE Signal Process. Lett.*, 2017, **24**, (3), pp. 279–283

[50] Zhang, H., McLoughlin, I., Song, Y.: 'Robust sound event recognition using convolutional neural networks'. Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Queensland, Australia, April 2015, pp. 559–563

[51] Choi, K., Fazekas, G., Sandler, M., *et al.*: 'Convolutional recurrent neural networks for music classification'. Proc. of the IEEE Int. Conf. on Acoustics, Speech and Signal Processing, New Orleans, USA, March 2017, pp. 2392–2396

[52] Young, S., Evermann, G., Gales, M., *et al.*: '*The HTK book*' (Cambridge University, England, 2002)

[53] Chang, C.C., Lin, C.J.: 'LIBSVM: a library for support vector machines', *ACM Trans. Intel. Syst. Technol.*, 2011, **2**, (3), p. 27

[54] Kaiser, J.: 'On a simple algorithm to calculate the 'energy' of a signal'. Proc. of the Int. Conf. on Acoustics, Speech and Signal Processing, New Mexico, USA, April 1990, pp. 381–384

[55] Allen, A.O.: '*Probability, statistics, and queueing theory with computer science applications*' (Academic Press Inc., Orlando, FL, USA, 1978)

[56] Schlosberg, H.: 'Three dimensions of emotion', *Psychol. Rev.*, 1954, **61**, (2), pp. 81–88