Speech Communication Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO

Leila Kerkeni, Youssef Serrestou, Kosai Raoof, Mohamed Mbarki, Mohamed Mahjoub, Catherine Cléder

Automatic Speech Emotion Recognition using an Optimal Combination of Features based on EMD-TKEO

Leila Kerkeni[a,b,*], Youssef Serrestou[a], Kosai Raoof[a], Mohamed Mbarki[c], Mohamed Ali Mahjoub[b], Catherine Cleder[d]

[a] *LAUM UMR CNRS 6613, Le Mans Université, France*
[b] *LATIS lab, ENISo Université de Sousse, Tunisia*
[c] *ISSAT, Université de Sousse, Tunisia*
[d] *CREN lab, Université de Nantes, France*

Abstract

In this paper, we propose a global approach for speech emotion recognition (SER) system using empirical mode decomposition (EMD). Its use is motivated by the fact that the EMD combined with the Teager-Kaiser Energy Operator (TKEO) gives an efficient time-frequency analysis of the non-stationary signals. In this method, each signal is decomposed using EMD into oscillating components called intrinsic mode functions (IMFs). TKEO is used for estimating the time-varying amplitude envelope and instantaneous frequency of a signal that is supposed to be Amplitude Modulation-Frequency Modulation (AM-FM) signal. A subset of the IMFs was selected and used to extract features from speech signal to recognize different emotions. The main contribution of our work is to extract novel features named modulation spectral (MS) features and modulation frequency features (MFF) based on AM-FM modulation model and combined them with cepstral features. It is believed that the combination of all features will improve the performance of the emotion recognition system. Furthermore, we examine the effect of feature selection on SER system performance. For classification task, Support Vecto Machine (SVM) and Recurrent Neural Networks (RNN) are used to distinguish seven basic emotions. Two databases- the Berlin corpus, and the Spanish corpus- are used for the experiments. The results evaluated on the Spanish emotional database, using RNN classifier and a combination of all features extracted from the IMFs enhances the performance of the SER system and achieving 91.16 % recognition rate. For the Berlin database, the combination of all features using SVM classifier has 86.22% recognition rate.

Keywords: Speech Emotion Recognition, EMD, TKEO, feature extraction, spectral modulation, feature selection, machine learning, RNN, SVM, classification, human-computer interaction.

*I am corresponding author
Email address: kerkeni.leila@gmail.com (Leila Kerkeni)

1. Introduction

A speech emotion recognition (SER) is an important issue in the development of human-computer interactions (HCI). It consists of predicting, from the speech signal, the emotion state of a speaker. There are many applications of detecting the emotion of the persons such as robot interface, audio surveillance, web-based E-learning, commercial applications, clinical studies, entertainment, banking, call centers, car board systems, computer games etc. For classroom orchestration or E-learning, information about the emotional state of students can provide a focus on enhancement of teaching quality. For example teacher can use SER to decide what subjects can be taught and must be able to develop strategies for managing emotions within the learning environment. That is why learners emotional state should be considered in the classroom. The performance of SER system depends on the quality of: features used to distinguish emotion, classifiers and data set used for training. For SER, several classification methods are used such as Support Vector Machines (SVM) [1], [2], [3], Gaussian Mixture Model (GMM) [4], Hidden Markov Model (HMM), Neural Networks (NN) [5], [4], [3], Recurrent Neural Networks (RNN) [6], [7], [8], and Linear Regression (LR) [9]. In this paper, we are particularly interested in the most important step on SER which is feature extraction. Many methods were used also at the stage of obtaining features from the signal. Among these approaches based on linearity and stationarity hypothesis is the Fast Fourier Transform (FFT) method which is the most useful method for frequency domain feature extraction. Nevertheless, this method loses some information [10], that may be useful for the classification task, in the time domain. To minimise this problem of stationarity, the Short Time Fourier Transform (STFT) is proposed to improve the traditional Fourier Transform [11], [10], [12]. The STFT consists in repeating the multiplication of a signal by a shifted short time windows and performing a Fourier Transform on the produced signal. Nonetheless, the STFT is also limited by the fundamental uncertainty principle, according to which time and frequency cannot simultaneously be resolved with the same precision. The issue of non-linearity remains problematic [13] and the STFT is also not appropriate to extract features from nonlinear systems. However, the speech signal is naturally non-stationary. Thus, it is more convenient to apply a non-stationary and nonlinear signal processing methods on the speech signal. The difficulties associated with an accurate recognition of emotion from speech remains challenging due to variability, complexity and subtle changes of non-linear features of the speech emotion. In order to overcome this problem, we used the EMD method, introduced by Huang et al. [14], to decompose and analyze the input signal which may be non linear and/or non stationary without losing its original properties. This method has been successfully applied to the area of speech emotion recognition (SER) [15], [16], [17], [18]. A review of the advancements in the non-conventional analysis of speech signals was written by Rajib Sharma et al. [19]. The EMD method decomposes the time series data into a number of components holding the highest local frequency, which is a collection of intrinsic mode functions (IMFs). However, the decomposition results using EMD method often suffer from mode-mixing problem [20], [21]. To solve this problem, Huang [20] proposes Ensemble empirical mode decomposition (EEMD). This method consists to add a Gaussian white noise to the original signal. In

the last few years, the EMD method combined with an energy operator is also used as an alternative technique that can improve the time-frequency analysis of signals [22]. Different energy operators were used for the changes in the sub-band signals. In [15] and [16], it is seen that effective results are obtained in the case of EMD and Teager-Kaiser Energy Operator (TKEO) being used together. TKEO can track the instantaneous amplitude and instantaneous frequency of the AM-FM component at any instant. The combination of EMD and Teager-Kaiser Energy Operator (TKEO) show an important role for time-frequency analysis and particularly for AM-FM signal demodulation [23]. Features extracted using the methods described above might contain redundant and irrelevant information which may affect SER system accuracy. Feature selection can solve this problem. Several methods for feature selection were found in literature and they are being widely used. The work by Tang and al. [24] summaries these methods. Examples of feature extraction techniques include Linear Discriminant Analysis (LDA), Principle Component Analysis (PCA), Sequential Forward Selection (SFS), Recursive Feature elimination (RFE). All these methods aim to remove the redundant and irrelevant features, so as to reduce the number of features and to improve model accuracy.

In this paper, we propose a global approach for SER. We presented a novel features named modulation spectral (MS) features and frequency modulation features (MFF) based on AM-FM demodulation and formant tracking. We combined them with the cepstral features: Energy Cepstral Coefficients (ECC), Frequency weighted Energy Cepstral Coefficients (EFCC) and the Mel Frequency Cepstral Coefficients based on the Reconstructed Signal (SMFCC). A method to select speech emotion feature named Recursive Feature Elimination (RFE) is applied after. The performance of the new features extracted was improved by speaker normalization. The Recurrent Neural Networks (RNN) and Support Vector Machines (SVM) models are used here to classify seven emotion status. In previous work [25], we compared several classifiers and the results of emotion classification by RNN, SVM and LR classifiers show that the RNN and SVM classifiers can improve classification accuracy significantly for both databases used in our work.

The next section of this paper gives an overview of the proposed emotion recognition system. The AM-FM modulation model using EMD and TKEO methods are explained in section 3. In section 4, we presented the features extracted in our work. Section 5 provides the feature selection method used in this paper. The RNN and SVM models used for classification are presented in section 6. The experimental results based on the features extracted in Section 4, are presented and discussed in Section 7. This section introduces also the databases employed. Finally, Section 8 concludes this work.

2. Overview of the emotion recognition system

Overall block diagram of our proposed system is shown in Figure 1. Among the most important step for developing a SER system is the extraction of speech features. For an efficient emotion recognition system, the best features are selected in this work to classify any speech signal into one of seven emotions. After that a machine learning algorithm is trained taking these features. Before features are being extracted from the input speech signal,

the following pre-processing is done to increase information content: i) decompose speech signal into IMFs using EMD sifting process ii) applying the nonlinear TKEO operator to the extracted IMFs. The process of each block is explained in detail in the following sections.

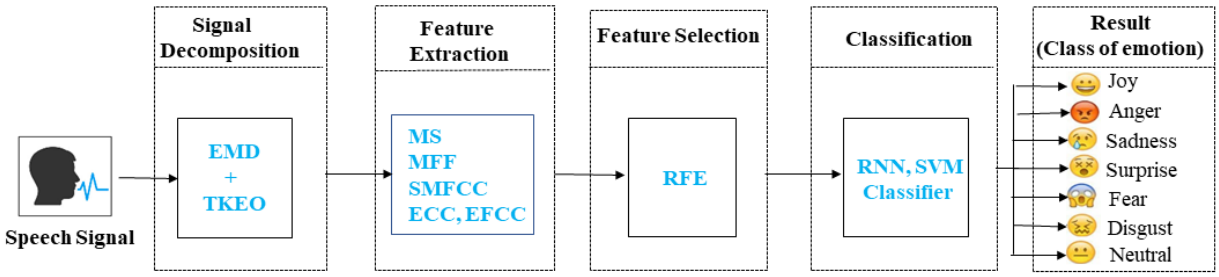


Figure 1: Block diagram of the proposed system.

3. AM-FM modulation model for speech analysis

In this paper, the AM-FM modulation model is used for speech analysis which represents the speech signal as the sum of formant resonance signals. Maragos and al. [26] describes this speech resonance $r_i(t)$ as a signal with a combined amplitude modulation (AM) and frequency modulation (FM) structure:

$$r_i(t) = Re\left(a_i(t)\, exp\left[\, j\ \phi_i(t)\right]\right) \tag{1}$$

According to this model, the speech signal can be expressed as follows:

$$x(t) = \sum_{i=1}^{N} r_i(t) + res(t) \tag{2}$$

where $res(t)$ is the last few components, which are low-frequency trend-like waveforms and it is excluded in general from the speech signal [19], Re represents the real part, $\phi_i(t)$ represents the phase, $a_i(t)$ is the instantaneous amplitude and $f_i(t)$ the instantaneous frequency of the ith IMF:

$$\phi_i(t) = 2\pi \int f_i(t)\mathrm{d}t$$

In [27], the speech resonance signal $r(t)$ is extracted from the speech signal $x(t)$ through bandpass filtering (Gabor filter). In our work, the EMD is used, which represents speech as the sum of IMFs ($r_i(t)$). Next, the TKEO is used to demodulate the resonance signals $r_i(t)$ into the amplitude envelope ($a_i(t)$) and instantaneous frequency ($f_i(t)$) signals. The following sections provide a detailed description.

3.1. Empirical Mode Decomposition (EMD)

The EMD is a method for decomposing any non-stationary signal into a collection of intrinsic mode functions (IMFs) which are mono-component AM-FM signals. The extraction of IMFs is non-linear, but their recombination for exact reconstruction of the signal is linear. In fact, the addition of all IMFs allows linear reconstruction of the original signal without loss or distortion of the initial information.

3.1.1. Intrinsic Mode Function (IMF)

A function is called an intrinsic mode function when it satisfies the following properties:

- The number of extrema (maxima+minima) in the signal must be equal or differ at most by one to the number of zero crossing;
- The mean of the envelopes defined by local maxima and local minima must be zero at all times.

3.1.2. Algorithm of the EMD

The EMD decomposition procedure for extracting an IMF is called the sifting process and is described as follows (Algorithm 1):

Algorithm 1 Empirical Mode Decomposition (EMD)

Input: A speech signal $x(t)$
Output: A collection of intrinsic mode functions (IMFs)

1. Compute all local extrema in the signal $x(t)$: the local maxima and local minima;
2. Construct the upper envelope $E_u(t)$ and lower envelope $E_l(t)$ by joining the local maxima and local minima with a cubic spline in the given signal $x(t)$;
3. Calculate the mean of the envelopes: $m(t) = (E_u(t) + E_l(t))/2$;
4. Subtract the mean from the original signal $x(t)$, then get a new data sequence $r(t)$ from which the low frequency is deleted: $r(t) = x(t) - m(t)$;
5. Repeat steps 1-4 until $r(t)$ is an IMF (that satisfy the two conditions above);
6. Subtract this IMF $r(t)$ from the original signal $x(t)$: $res(t) = x(t) - r(t)$;
7. Repeat steps 1-6 till there are no more IMFs left in the residual signal $res(t)$ in order to obtain all the IMFs $r_1(t)$, $r_2(t)$,..., $r_N(t)$ of the signal $x(t)$.

The process terminates when the residual $res(t)$ is either a constant, a monotonic slope or a function with only one extrema. The result of the EMD process produces N IMFs ($r_1(t)$, $r_2(t)$,..., $r_N(t)$) and residue signal ($res_N(t)$). Consequently, the original data sequence $x(t)$

may represent the sum of a group of IMF values plus a residual term as shown in the following formula:

$$x(t) = \sum_{i=1}^{N} r_i(t) + res_N(t) \tag{3}$$

In this method, each input signal is decomposed into a finite number of IMFs using the previously described EMD algorithm. Each IMF can be then analyzed separately in order to obtain features for emotion classification. In [28], authors are limited to the first five IMFs providing that these IMFs give a sufficient information about of energy and pitch in their cases. Several number of IMFs are used in our work of iterations to determine the optimal number of IMFS for reconstructing the signal without any loss of important information for each database. We have seen that after a certain number, the IMFs do not provide information.

3.2. Teager-Kaiser Energy Operator (TKEO)

The IMFs $r(t)$ obtained from EMD as such do not convey any information nor provide a meaning. Applying the Teager-Kaiser Energy Operator (TKEO) in the IMFs allows an estimate of time-varying amplitude envelope and instantaneous frequency, which provides some physical meaning. The Teager-Kaiser energy operator is a nonlinear operator that calculates the energy of mono-component signals as the product of the square of the amplitude and the frequency of the signal. The instantaneous characteristics of these signals are then obtained by the application of the Discrete Energy Separation Algorithm (DESA-2) [26]. It will be shown that TKEO improves the estimation of instantaneous characteristics of the vibration data compared to other commonly used techniques, e.g. the Hilbert Transform. A method based on EMD algorithm of Huang et al. [29] and TKEO is called Teager-Huang Transform (THT). TKEO is defined for an IMF $r_i(t)$ in a continuous time as [30]:

$$\Psi[r_i(t)] = [\dot{r}_i(t)]^2 - r_i(t)\ddot{r}_i(t) \tag{4}$$

where $\dot{r}_i(t)$ and $\ddot{r}_i(t)$ are respectively the first and the second time derivatives of $r_i(t)$. For a discrete time signal $r_i(n)$, the time derivatives of the equation 4 can be proposed as [31]:

$$\Psi[r_i(n)] = r_i^2(n) - r_i(n+1)r_i(n-1) \tag{5}$$

where n is the discrete time index.

The following equations describe exactly the instantaneous frequency $f(n)$ and instantaneous amplitude $a(n)$ at any time instant of the IMF $r_i(n)$ [26]:

$$f(n) = \frac{1}{2}\arccos(1 - \frac{\Psi[x(n+1) - x(n-1)]}{2\Psi[x(n)]} \tag{6}$$

$$\mid a(n) \mid = \frac{2\Psi[x(n)]}{\sqrt{\Psi[x(n+1) - x(n-1)]}} \tag{7}$$

Researchers in [32] found that the energy operator approximation incurs a high-frequency error component. So they proposed to eliminate the high-frequency error component by

filtering the energy operator output through an appropriate low-pass filter as shown in Figure 2 . For this we used in our work a seven-point linear binomial smoothing filter with impulse response (1, 6, 15, 20, 15, 6, 1) [32].

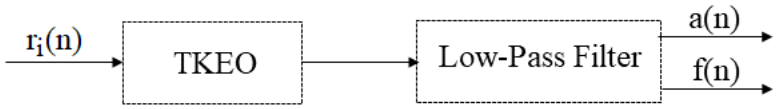


Figure 2: The block diagram of the smoothed energy operator.

4. Feature Extraction

The crucial process in SER system consists of extracting the essential information from the signal. In this study, we have used EMD combined with the TKEO to extract features as illustrated in Figure 3, 4, 5 and 8. Each set of features is explained and detailed in the following sections.

4.1. Cepstral features

The cepstral features are important parameters in SER, which exploites the property of human auditory system and there are derived from the signal spectrum. In this section, we extracted these features to combine them with the new features presented in Section 4.2. Both types of features contain emotion information, it is believed that the combination of them will improve the performance of the emotion recognition system.

4.1.1. SMFCC features

MFCC is widely used for speech emotion recognition [2]. It has good performance in description of the human ear's auditory characteristics. Assuming that a speech signal is a short-term stationary process, MFCC convey the signal's short term spectral only and omit some important information. But really, the speech signal posses complex and random changes and the existence of signal trend would result in great error with respect to the power spectral analysis in frequency domain or correlation analysis in time domain, and even lead to the completely loss of the authenticity of the low-frequency spectrum. Hence the need to remove this signal trend. In this section, we calculate the features extracted after the removal of signal trend $T(n)$. These features, called SMFCC and introduced in [16], provide a more accurate description of the distribution of energy in the frequency domain. The EMD based signal reconstruction method is carried out to conduct the SMFCC feature extraction. The process of extracting SMFCC is represented in Figure 3. Firstly, the EMD method is conducted on the input speech signal. Secondly, the signal trend, according to [15], is calculated by the formula below (equation 9) and removed after from the original signal using the zero-crossing rate (ZCR) detection method. $T(n)$ is the sum of IMFs $r(n)$ that respect the following constraint [15]:

$$\frac{R_{r_i}}{R_{r_1}} < 0.01 \quad (i = 2, ...n) \tag{8}$$

where R represents the zero-crossing rate.

$$T(n) = \sum_i r_i(n) \tag{9}$$

Subsequently, the final signal $S(n)$ is obtained by the subtraction of the signal trend $T(n)$ from the original speech signal $x(n)$ [15].

$$S(n) = x(n) - T(n) \tag{10}$$

The SMFCC is extracted from the obtained reconstituted signal $S(n)$ through FFT algorithm and discrete Cosine Transform (DCT).

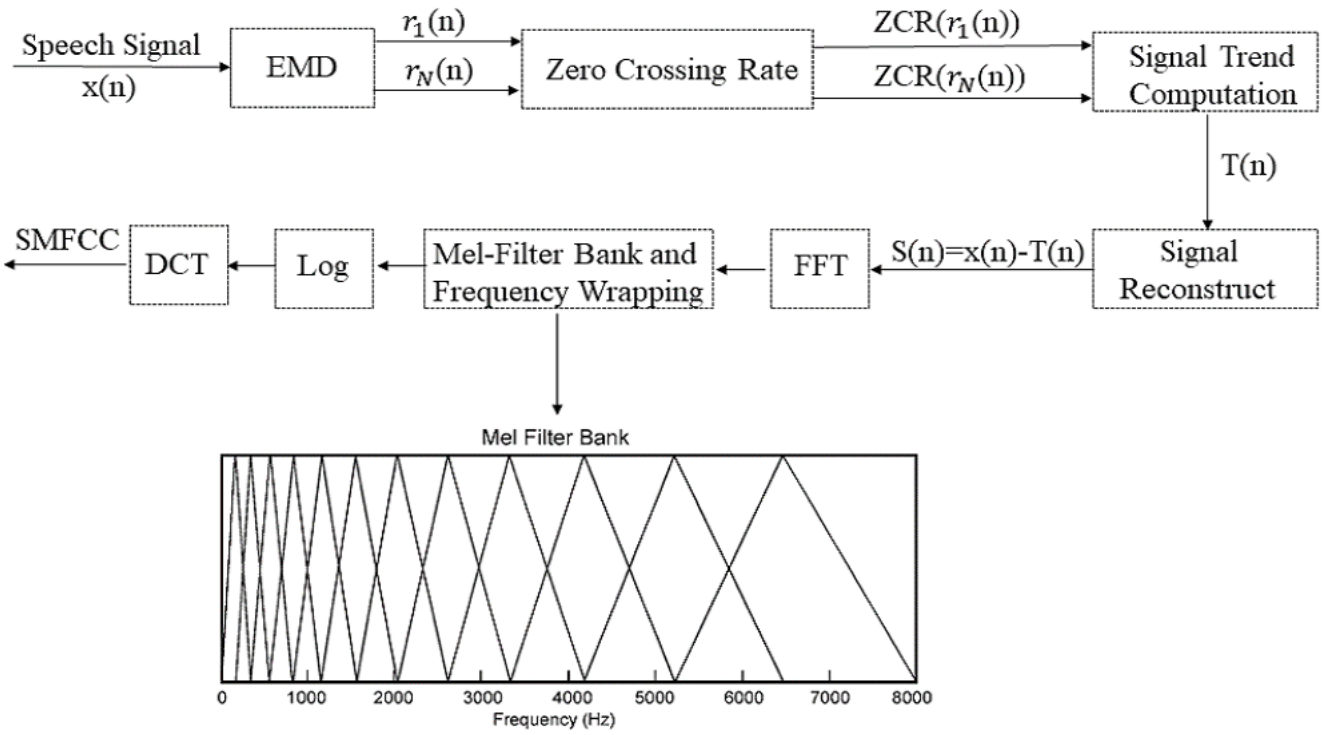


Figure 3: Schema of SMFCC extraction.

4.1.2. Energy Cepstral Coefficients (ECC) and Frequency weighted Energy Cepstral Coefficients (EFCC) features

In [19], authors demonstrate that the distribution of spectral energy varies under different emotions. This means that emotion can affect energy distribution of speech among different frequency bands. Hence the importance of extracting the ECC feature that provides the distribution of energy in the Hilbert spectrum (in the time domain). In standard approach, the instantaneous energy of each IMF is considered to be proportional with the amplitude and has nothing to do with the instantaneous frequency. According to the physical model

of instantaneous frequency, not only the instantaneous energy is included in the energy envelope of $a(t)$, but also in the instantaneous frequency $f(t)$. Since the internal energy is proportional with its circumferential velocity by physical sense [19]. On the basis of the above considerations, the researchers developed the instantaneous frequency-weighted energy (EFCC) to improve the ECC features presented above. The standard implementation of computing the Energy Cepstral Coefficients (ECC) and the Frequency weighted Energy Cepstral Coefficients (EFCC) [15] is shown in Figure 4. The first processing step is the decomposition of speech signal into IMFs using EMD. The imf's instantaneous amplitude ($a(i,n)$) and the IMF's instantaneous frequency ($f(i,n)$) are estimated using TKEO at the second step. The third processing step tries to register the instantaneous amplitude and frequency into short and overlapping frames (the duration of frame is 250 ms and the overlapping is equal to 64 ms) which gives $a_k(i,n)$ and $f_k(i,n)$. The fourth processing step is the computation of the Marginal Hilbert Spectrum (MHS) using Hilbert spectrum. It offers a measure of the total amplitude from each frequency. Therefore, the spectrum is decomposed into different frequency bands (12 bands) and the power of each frequency band is computed. Afterwards, the natural logarithm of sub-band energy and complement the discrete cosine transform (DCT) are computed. The first 12 DCT coefficients provided the ECC and EFCC values used in the classification process.

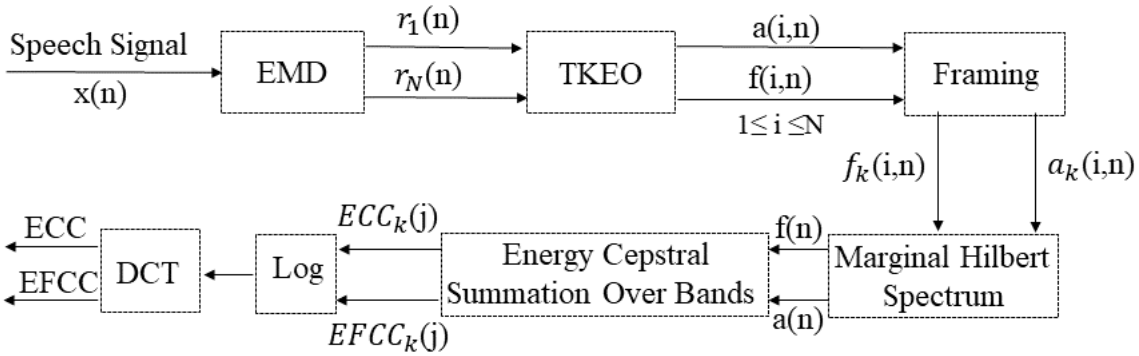


Figure 4: Schema of ECC and EFCC extraction.

The MHS $h_j(f)$ gives the probability that an ascillation of frequency could have occured locally at some time during the entire duration of the signal [19] and it is obtained by:

$$h_j(f) = \sum_{n=1}^{L_f} H(f,n)\mathbb{1}_{B_j}(f) \tag{11}$$

where L_f represents the length of frame in samples and $H(f,n)$ is the Hilbert spectrum. It is defined as the instantaneous energy envelope in a time-frequency space, which is the the squared magnitude of the amplitude envelope [19]. $H(f,n)$ is derived by:

$$H(f,n) = \sum_{i=1}^{N} a(i,n)^2 \mathbb{1}_{\{f(i,n)\}}(f) \tag{12}$$

where i represents a particular IMF, and B_j represents a particular sub-band. B_j is defined as: $B_j = [f_c(j-1), f_c(j+1)]$ where f_c represents the frequency center.

The indicator function of a subset Ω of a set x is defined as:

$$\mathbb{1}_{\Omega}(x) = \begin{cases} 1 & \text{if } x \in \Omega \\ 0 & \text{if } x \notin \Omega \end{cases}$$

4.1.2.1 Energy Cepstral Coefficients (ECC)

ECC is given in a continuous time by [19]:

$$ECC(B_j, \Pi_i) = \int_{f \in B_j} h_j(f) \mathrm{d}f, \ t \in \Pi_i \ , \ j = 1, ..., 12 \tag{13}$$

where B_j represents a particular sub-band, and Π_i represents a particular speech frame. In this work, these features are calculated, by using a discrete time of speech signal, as follows:

$$ECC_k(j) = \sum_{i=1}^{N} \frac{1}{L_F} \sum_{n=1}^{L_F} a_k^2(i,n) \mathbb{1}_{B_j}(f_k(i,n)), \ j = 1, ..., 12 \tag{14}$$

4.1.2.2 Frequency weighted Energy Cepstral Coefficients (EFCC)

EFCC is given in a continuous time by [19]:

$$EFCC(B_j, \Pi_i) = \int_{f \in B_j} f(t) h_j(f) \mathrm{d}f, \ t \in \Pi_i \ , \ j = 1, ..., 12 \tag{15}$$

Where B_i represents a particular sub-band, and Π_i represents a particular speech frame. For a discrete case, EFCC are calculated as follows:

$$EFCC_k(j) = \sum_{i=1}^{N} \frac{1}{L_F} \sum_{n=1}^{L_F} f_k(i,n) a_k^2(i,n) \mathbb{1}_{B_j}(f_k(i,n)), \ j = 1, ..., 12 \tag{16}$$

where B_j represents a particular sub-band, i represents a particular IMF, and k represents a particular speech frame. In our work, we extract the first 12-order of the ECC and EFCC coefficients where the speech signals are sampled at 16 KHz. For each order coefficients, we compute the mean, variance, coefficient of variation, Kurtosis and Skewness. Each ECC and EFCC feature vector is composed of 60 points.

4.2. AM-FM modulation features

In this section, we presented our new features based on AM-FM modulation model. This model captures both long-term modulation features and modulation frequency features, thereby conveying information that is important for human speech perception.

4.2.1. Modulation Spectral (MS) features

The modulation spectral features (MSF) introduced in [33] are extracted specifically to solve the problems of short-term spectral features (MFCC) and to better model the nature of human auditory perception. The method is based on emulating the Spectro-temporal (ST) processing performed in the human auditory system and considers regular acoustic frequency jointly with modulation frequency. These features are based on decomposition of speech signal by an auditory filter bank and the computation of the Hilbert envelope of each band. The Hilbert envelope is based on the Fourier transform (FT) and hence some of the limitations of FT are also associated with it. For this reason we propose in this work a new way of extracting these features based on speech analysis using AM-FM modulation model. The steps for extracting modulation spectral (MS) features depicted in Figure 5. After applying TKEO on IMFs decomposed by means of EMD. A modulation filter bank is further applied to the instantaneous amplitude to perform frequency analysis. The spectral contents of the modulation signals are referred to as modulation spectra, and the proposed features are named modulation spectral (MS) features. The energy, taken over all frames in every spectral band provides a feature ($E(i,j)$).

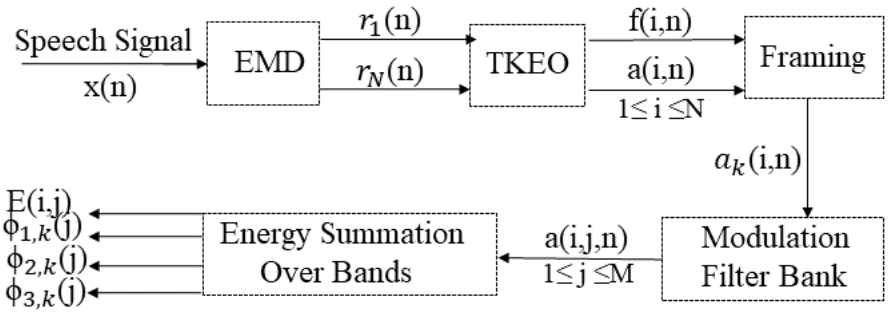


Figure 5: Schema of MS features extraction.

Energy in every spectral band, is defined as:

$$E(i,j) = \sum_{k=1}^{N_f} E_k(i,j) \tag{17}$$

Where $E_k(i,j)$ is the Energy over bands, N_f is the number of frame for $1 \leq j \leq 8$. For each frame k, $E(i,j)$ is normalized to unit energy before further computation:

$$\sum_{i,j} E_k(i,j) = 1$$

Three spectral measures Φ_1, Φ_2, Φ_3 are then calculated on a per-frame basis [33]. The first spectral measure is the spectral energy $\Phi_{1,k}(j)$ which is defined as the mean of the energy samples belonging to the jth modulation band ($1 \leq j \leq 8$):

$$\Phi_{1,k}(j) = \frac{\sum_{i=1}^{N} E_k(i,j)}{N} \tag{18}$$

For frame k, $\Phi_{2,k}(j)$ is the spectral flatness which is defined as the ratio of the geometric mean of Φ_1 to the arithmetic mean. Φ_2 is thus defined as follows:

$$\Phi_{2,k}(j) = \frac{\sqrt[N]{\prod_{i=1}^{N} E_k(i,j)}}{\Phi_{1,k}(j)} \tag{19}$$

In our work, Φ_2 is expressed on a logarithmic scale as follows:

$$\log \Phi_{2,k}(j) = \frac{1}{N} \sum_{i=1}^{N} \log E_k(i,j) - \log \Phi_{1,k}(j)$$

The last spectral measure employed Φ_3 is the spectral centroid which provides a measure of the "center of mass" of the spectrum in each modulation band. For jth modulation band, Φ_3 is defined as:

$$\Phi_{3,k}(j) = \frac{\sum_{i=1}^{N} f(i) E_k(i,j)}{\sum_{i=1}^{N} E_k(i,j)} \tag{20}$$

Where $f(i) = i$. In [33], the authors experiment two similar types of frequency measure $f(i)$. It being in our case the index of the ith critical-band filter, i.e., $f(i) = i$. We choose it for its simplicity. We have noted that there is a considerable correlation between two adjacent modulation bands. In order to reduce the high correlation for the spectral flatness and the centroid parameters, $\Phi_{2,k}(j)$ and $\Phi_{3,k}(j)$ are only computed for $j \in \{1, 3, 5, 7, 8\}$.

Various statistics, mean, variance, coefficient of variation, kurtosis and skewness, are extracted from energy in each spectral band. Along with those statistics, the mean and variance of the spectral energy, spectral flatness and spectral centroid are evaluated, and used as features.

Figure 6 shows the mean of Energy $E(i,j)$ for the seven emotions (*anger*, *boredom*, *disgust*, $fear$, *happiness*, *sadness* and *neutral*) in the Berlin database. Every $E(i,j)$ represents the mean over all frames from all speakers for an emotion. It is seen from this Figure that the energy distribution over the joint IMF-modulation frequency plane is similar for some emotions that could become confusion pairs, like anger and happiness, boredom and neutral. But it is very distinct for some others emotions, like *anger* and *sadness*, they could be well discriminated from each other.

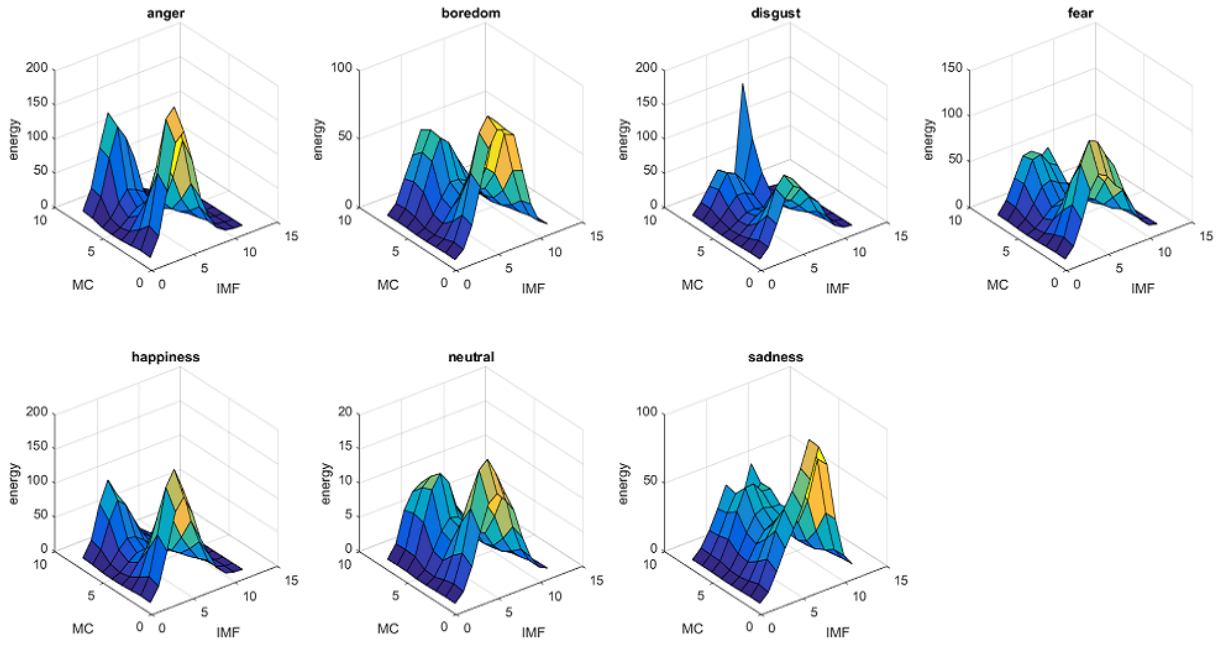


Figure 6: Mean $E(i,j)$ for seven emotion; For each emotion, $E_k(i,j)$ is averaged over all frames from all speakers of that emotion; "MC" denotes the modulation frequency channels.

Figure 7 shown that the emotions *sadness* and *neutral* have significantly more low acoustic frequency energy than anger emotion, which is approximately the same in the previous work [33]. For the emotion of *anger* the dispersion is more greater. However, the less expressive emotions such as *sadness* exhibit more prominently low pass modulation spectral shapes, which suggest of lower speaking rates.

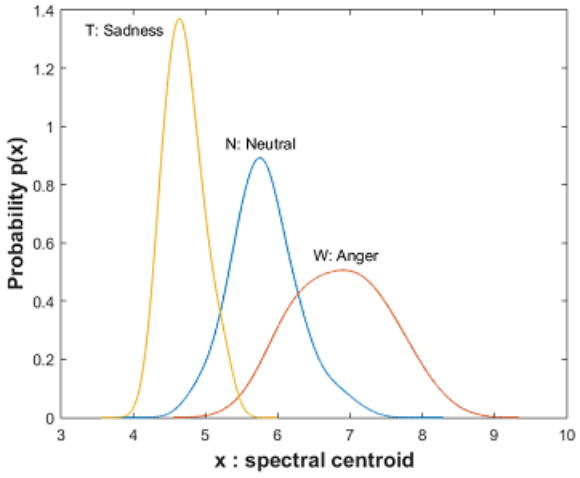


Figure 7: Estimated Probability Density Function of $\overline{\Phi}_3$ for three basic emotions (*sadness*, *neutral* and *anger*).

4.2.2. Modulation Frequency Features (MFF)

Studies on speech perception demonstrated that the most important perceptual information lies at low modulation frequencies [34]. In this section, we will illustrate how modulation frequency analysis can be formulated and applied to the problem of SER. The AM-FM modulation model is used for the analysis of signal and we are interested in the frequency distribution of the energy of the speech signal, in a frequency band. The steps to compute the frequencies based on AM-FM decomposition of the input speech are shown in Figure 5. The speech signal is decomposed into several intrinsic mode functions (IMFs) by EMD. The instantaneous amplitude envelope and frequency function of each IMF is estimated using the TKEO. Then, they are used to obtain short-time estimates of mean instantaneous frequency and bandwidth over a frame.

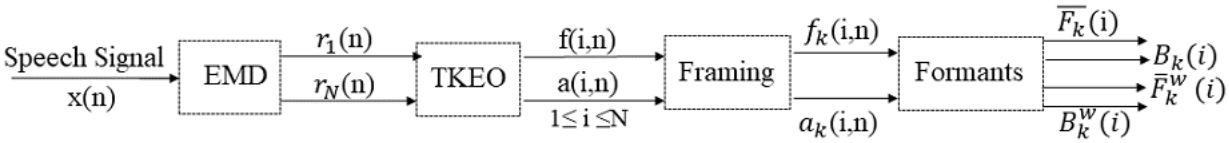


Figure 8: Schema of MFF extraction.

The mean amplitude weighted instantaneous frequency (F_w) and the mean amplitude weighted instantaneous band-width (B_w) are described for a continuous time signal in [19] by :

$$F_w = \frac{\int_{t_0}^{t_0+T} f(t)a^2(t)\mathrm{d}t}{\int_{t_0}^{t_0+T} a^2(t)\mathrm{d}t} \tag{21}$$

$$B_w^2 = \frac{\int_{t_0}^{t_0+T} [\{\dot{a}(t)/2\pi\}^2 + \{f(t) - F_w\}^2 a^2(t)]\mathrm{d}t}{\int_{t_0}^{t_0+T} a^2(t)\mathrm{d}t} \tag{22}$$

where t_0 and T represent the beginning and duration of analyzed frame, $f(t)$ and $a(t)$ represent respectively the instantaneous frequency and amplitude envelope of each AM-FM signal and $\dot{a}(n)$ is discretized and defined by:

$$\dot{a}(t) = a(t+1) - a(t)$$

Using the amplitude envelope and instantaneous frequency signals short-time estimates are proposed in this work such as: the mean instantaneous frequency ($\overline{F}$), the mean instantaneous band width (B), the mean amplitude weighted instantaneous frequency ($\overline{F}_w$) and the mean amplitude weighted instantaneous band-width (B_w), are given in the discrete case by:

$$\overline{F}_k(i) = \frac{1}{L_F} \sum_{n=1}^{L_F} f_i(k,n) \tag{23}$$

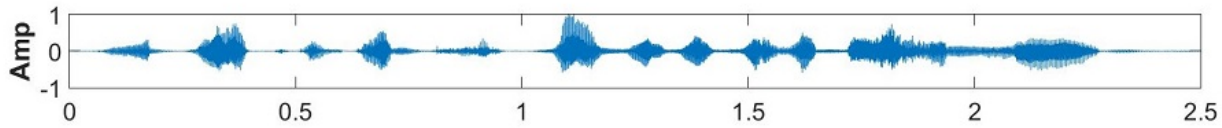
$$B_k(i) = \sqrt{\frac{1}{L_F}\sum_{n=1}^{L_F}(f_k(i,n) - \overline{F}_k(i))} \tag{24}$$

$$\overline{F}_k^w(i) = \frac{\sum_{n=1}^{L_F} f_k(i,n)a_k^2(i,n)}{\sum_{n=1}^{L_F} a_k^2(i,n)} \tag{25}$$

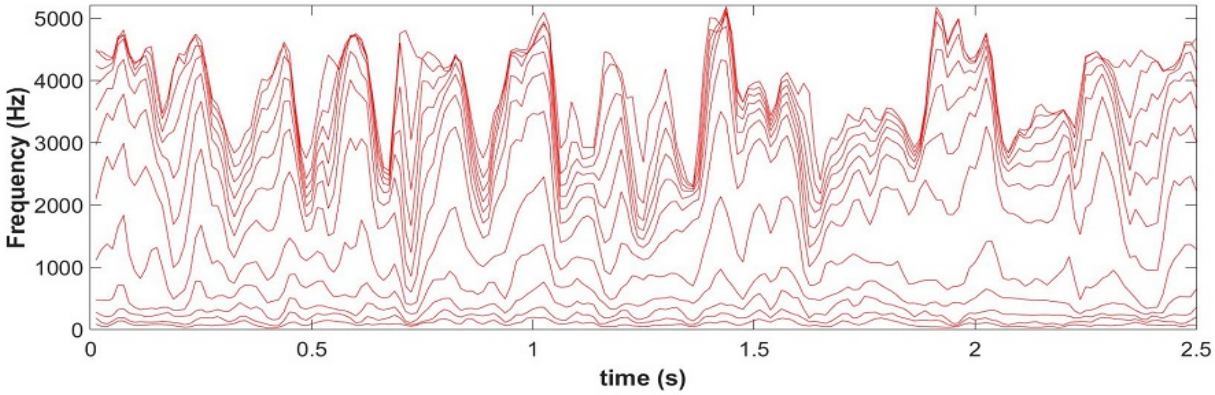
$$B_k^w(i) = \sqrt{\frac{\sum_{n=1}^{L_F}\{\dot{a}_k(i,n)/2\pi\}^2 + \{f_k(i,n) - \overline{F}_k^w(i)\}^2 a_k^2(i,n)}{\sum_{n=1}^{L_F} a_k^2(i,n)}} \tag{26}$$

where i represents a particular IMF, L_F represents a number of samples per frame, f_k and a_k represent respectively the instantaneous frequency and amplitude frequency at a particular speech frame k. In this work the mean, maximum and minimum of these two values (B, B_w) and the mean of ($\overline{F}$, $\overline{F}_w$) were computed for each feature and it was used for classification.

We can see the usefulness of AM-FM analysis from the time-frequency representation of a speech signal shown in Figures 9 and 10. For each IMF component, a short time frequency estimate $\overline{F}_k(i)$ is obtained at every frame using Equation 23. The time duration of the analysed frame is 25 ms with a frame-shift of 10 ms. The figures below indicate that AM-FM analysis could provide useful information to discriminate emotion in speech signal. They identify regions with dense clusters for the purpose of tracking formant frequencies and their bandwidths [19]. It may be noted from Figures 9 and 10 that the dense clusters of curves in the time-frequency representation are different between the "anger" speech signal and "neutral" speech signal, the difference is due to emotion. The estimated density of probability, in Figure 12, confirm this assumption.

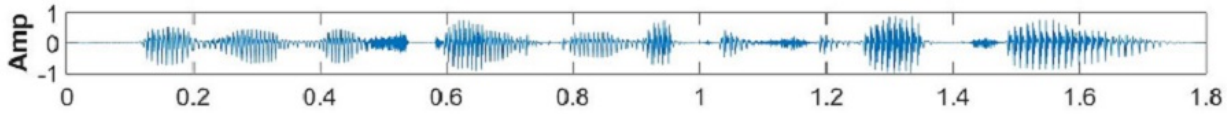


(a)

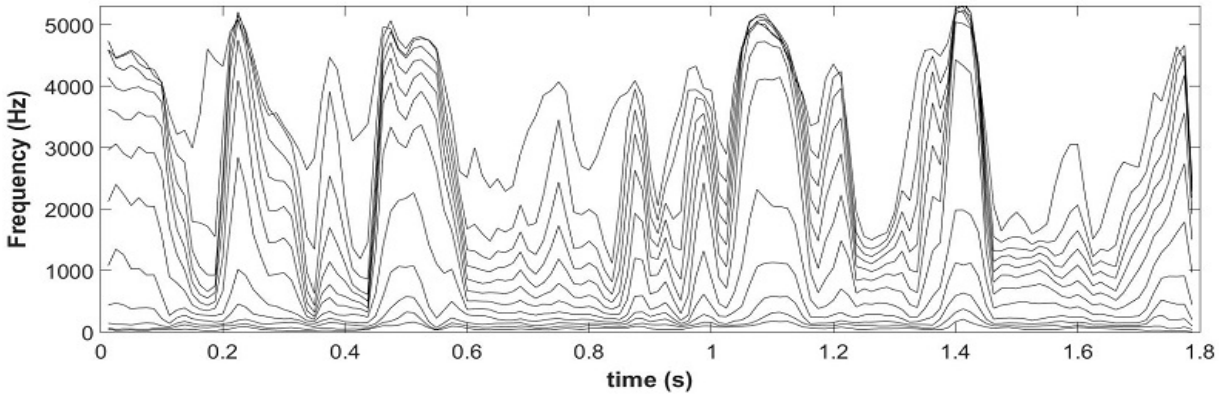


(b)

Figure 9: (9a) Signal of a "anger" speech file; (9b) The time-frequency representation of (9a) using 12 IMFs. Framesize of 25 ms with framshift of 10 ms is used.



(a)



(b)

Figure 10: (10a) Signal of a "neutral" speech file; (10b) The time-frequency representation of (10a) using 12 IMFs. Framesize of 25 ms with framshift of 10 ms is used.

In the case of 535 utterances here considered, we obtained that the fundamental frequency was embedded in the IMF10- IMF12. Figure 11 shown the mean values of the instantaneous frequency ($\overline{F}$), for indexes from 1 to 535, corresponding to each file on the Berlin database.

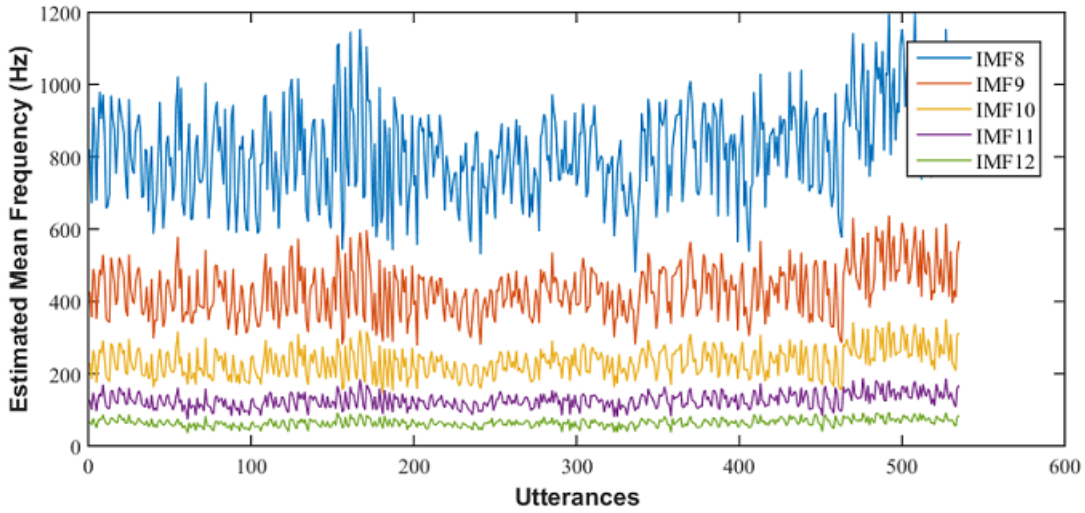


Figure 11: Estimated Mean Frequency from all speakers

In Figure 12, we find almost the same results as Figure 7. The *anger* emotion is shifting to the high frequencies (between 600 and 1300 Hz). However, the emotion of *sadness* towards the low frequencies.

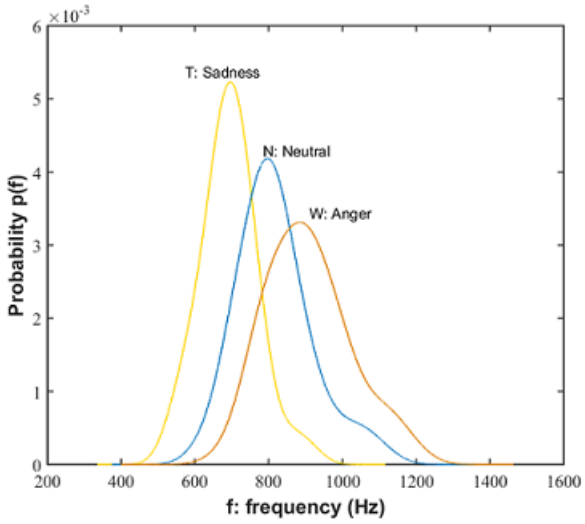


Figure 12: Estimated Probability Density Function of Mean Frequency for three basic emotions (*sadness*, *neutral* and *anger*).

5. Feature Selection (FS)

After extracting features from the speech signal, the most important thing is that how to select features that are capable of discriminating samples that belong to different classes. As reported by Aha and Bankert [35], FS is a process to select a subset of relevant features used to characterize a dataset so as to improve a learning algorithm's performance on a given task. The extracted features contain unneeded, irrelevant and redundant data that do not contribute to the accuracy of a predictive model or may in fact decrease the accuracy of the model. The extraction of features is not enough for a better performance of the classifier. So it is necessary to select features. FS allows not only to lead to higher recognition accuracy but also to reduce drastically the running time of the learning algorithms. In this study, an effective and a simple feature selection method named Recursive Feature Elimination (RFE) is used.

5.1. Recursive Feature Elimination (RFE)

RFE uses a model (eg. linear Regression or SVM) to select either the best or worst-performing feature, and then excludes this feature. These estimators assign weights to features (e.g., the coefficients of a linear model), so the goal of recursive feature elimination (RFE) is to select features by recursively considering smaller and smaller sets of features. First, the estimator is trained on the initial set of features and the predictive power of each feature is measured [36]. Then, the least important features are removed from current set of features. This procedure is recursively repeated on the pruned set until the desired number of features to select is eventually reached. In this work, we implemented the Recursive Feature Elimination method of feature ranking via the use of basic Linear Regression (LR-RFE) [37]. SVM-RFE is an SVM-based feature selection algorithm created by [38]. Using SVM-RFE, Guyon et al. selected an important feature sets. In addition to improving the classification accuracy rate, it can reduce classification computational time.

6. Classification models

In machine learning, classification is a supervised learning approach in which the computer program learns a mapping from the training samples in certain features space to some labels. This training samples may simply be bi-class (i.e identifying whether the person is male or female) or it may be multi-class too. There are many types of classification algorithms in machine learning: Linear Regression, Decision Tree, SVM, Naive Bayes, KNN, Random Forest, Neural Networks, etc. These algorithms can be applied to almost any data problem. Here we have used SVM and RNN classifiers.

6.1. Support Vector Machine (SVM)

SVM is a simple and optimal classifier in machine learning. The advantage of SVM compared to other classifiers is, for a limited training data it shows better classification performance [4]. SVM has been used extensively in many studies that related to audio emotion recognition and has been shown to give best performance in many SER works [1]

and [39]. It is built by mapping the training patterns into a higher dimensional feature space where the points can be separated using a hyper plane. The idea behind the SVM is to transform the original input set to a high dimensional feature space by using kernel function. The kernel make the SVM method much more flexible and powerful. SVM kernels include: linear, polynomial, radial basis function (RBF) and sigmoid [40]. SVM theoretical background can be found in [41]. A MATLAB toolbox implementing SVM is freely available in [42]. Here in our work we have used polynomial kernel [40], which is defined as:

$$K(x_i, x_j) = (\langle x_i, x_j \rangle + c_0)^d \tag{27}$$

where d is the degree of the polynomial. In general, nonzero c_0 is preferred [40].

6.2. Recurrent Neural Networks (RNNs)

RNN is recommended for sequence classification where there are some sequence of inputs over space or time and the task is to predict a category for the sequence. In recent years, RNN have demonstrated ground-breaking performance for sequence analysis [6], [43]. This work used the long short term memory (LSTM), which is a special kind of RNN capable of learning long-term dependencies, to implement the RNN layer. LSTM were developped by Hochreiter et al [44]. RNN models suffer from the vanishing gradient problem which increases with the length of the training sequences. LSTM solving this problem in classical RNN by using memory cells that allow information to accumulate during operation and uses feedback to remember previous network call states [45].

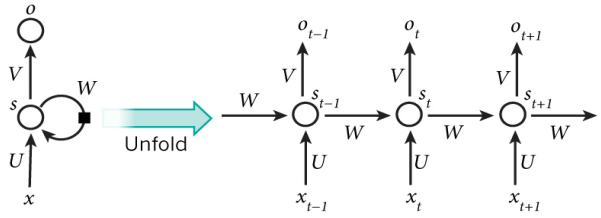


Figure 13: A basic concept of RNN and unfolding in time of the computation involved in its forward computation [6].

Figure 13 shows a basic concept of RNN implementation. Unlike traditional neural network that uses different parameters at each layer, the RNN shares the same parameters (U,V and W in Figure 13) across all steps. The hidden state formulas and variables are as follows:

$$s_t = f(Ux_t + Ws_{t-1}) \tag{28}$$

with:

- x_t, s_t and o_t are respectively the input, the hidden state and the output at time step t;
- U, V, W are parameters matrices.

7. Experimental results and analysis

7.1. Emotional databases

In the design of a SER system, one important thing upon which the performance of emotion recognition system depends is the database used to build its representation of humans emotions. Different types of databases have been used for the study of emotions in speech [46]. First type is the simulated databases, also called acted databases. Speakers are generally experienced professional actors. The actors are asked to express some sentences in different emotions. Another type is the induced databases, also called elicited databases. It is collected by provoking an emotional reaction using an artificial emotional situation. Emotional states are induced using different methods, such as videos, images, stories or computer games. Without the knowledge of the speaker that they have been recording, the databases are more natural and realistic than the simulated databases. The third type is the natural emotions. These emotions are obtained by recording speakers in natural situations such as call center conversations, TV talk shows, patient doctor conversations and so on. This type of material might obscure the exact nature of recorded emotions because of background noise, overlapping voices, artifacts, etc, which makes the quality of speech unsatisfactory. In this work we used acted databases because it is the only one that allows a complete control over the recorded text [47].

7.2. Berlin database

The Berlin database [48] is a German acted database. It is the most commonly used database in the context of speech emotion recognition, and also one of the common public databases which can be freely accessed by researches. It consists of recording 10 German sentences by 10 actors (5 male, 5 female). The final database consists of 535 utterances recorded in *anger*, *boredom*, *disgust*, *fear*, *happiness*, *sadness* and *neutral*. The samples for each emotion are distributed as shown in Figure 14.

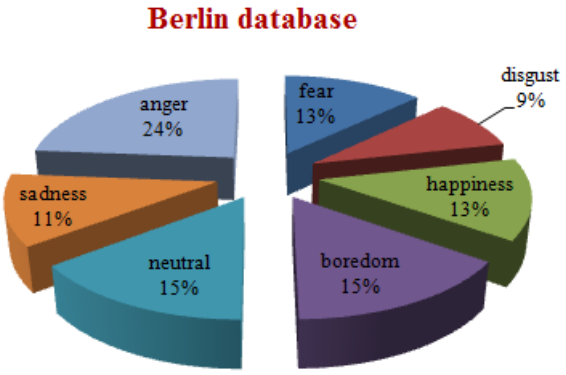


Figure 14: Emotions percentage distribution for the Berlin database.

7.3. Spanish database

The Spanish emotional speech database is available for researches use [49], and it contains more data (6041 utterances) compared to other created databases. Six basic or primary emotions (*anger*, *boredom*, *disgust*, $fear$, *happiness* and *sadness*) in addition to a *neutral* variations (normal, soft, loud, slow and fast) speaking style were recorded. The database contains two emotional speech recording sessions played by two actors (1 female, 1 male). The speakers had to portray the different emotions as shown in Figure 15.

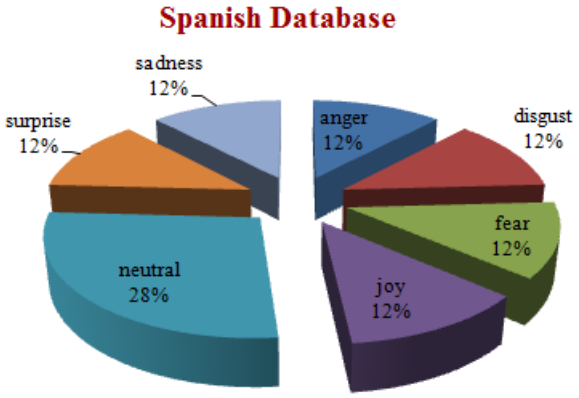


Figure 15: Emotions percentage distribution for the Spanish database.

7.4. Methodology, Results and analysis

For our experiments we used the two databases (Berlin and Spanish) described in section 7.1. All classification results are obtained under 10-fold cross-validation. Cross-validation is a common practice used in performance analysis that randomly partitions the data into N complementary subsets, with $N-1$ of them used for training in each validation and the remaining one used for testing. The system described in section 2 (Figure 1) was implemented with the different features presented in section 4. The impact of these features for the SER system accuracy is demonstrated in this section. Several number of IMFs are used in our work with regard to determine the optimal number of IMFs for reconstructing the signal without any loss of important information for each database. We have seen that after a certain number of iterations, the IMFs do not provide information. That is why the signal in our work is reconstructed from only the first nine modes (9 IMFs) for Spanish database and the first twelve modes (12 IMFs) for the Berlin database. The difference in the number of IMFs is due to the nature of the languages in the two databases. These IMFs was used after for feature extraction to recognize emotions. The neural network structure used is a simple LSTM. It consists of two consecutive LSTM layers with hyperbolic tangent activation followed by two classification dense layers. More detailed diagrams are shown in our previous work [25]. For computing kernel and related parameters of the SVM classifier, we used an automatic technique of choice (cross validation) with different kernel, epsilon and coefficient c. Features from data are scaled to $[-1, 1]$ before applying classifiers. Scaling features before recognition is important, because when a learning phase is fit on unscaled

data, it is possible for large inputs to slow down the learning and convergence, in some cases prevents the used classifier from effectively learning for the classification problem. The effect of speaker normalization (SN) step prior to recognition is investigated, and there are three different SN schemes that are defined in [40]. SN is useful to compensate for the variations due to speaker diversity rather than change of emotional state. We used in this section the SN scheme that has given the best results in [40]. The features of each speaker are normalized with a zero mean and a standard deviation of 1. It is well known that feature selection has a huge influence on SER systems [50]. A good set of features can facilitate and improve model learning. Therefore, here, we assess the impact of feature selection on our proposed system. Table 1 gives a description of the number of the features extracted corresponding to each feature extraction label for both databases. The difference in the number of features (MS, MFF) for two databases is related to the number of IMFs (12 and 9 IMFs respectively for Berlin and Spanish database).

Table 1: Description of the number of features extracted.

Label feature extraction	**Number of features**	
	Berlin DB	**Spanish DB**
MS	132	72
MFF	96	72
ECC	60	60
EFCC	60	60
SMFCC	60	60

The detailed numerical results were illustrated in Table 2 to Table 8. Tables 2 and 3 present the recognition rates of different features for both databases. These experiments using feature set without feature selection. As shown in Table 2, SVM classifier yields better results above 76.28 % with SMFCC feature for Berlin database. We obtain the same results obtained in [15]. These results improved in our work by speaker normalization and 81.13 % accuracy is achieved. From Table 1, it can be also concluded that applying SN improves recognition results for Berlin database. But this is not the case for the Spanish database, as demonstrated in Table 3. The same results are obtained with both classifiers. This can be explained by the number of speakers in each database, where the Berlin database contains 10 different speakers, compared to the Spanish database that contains only two speakers.

Table 2: Recognition rates for different features on Berlin database; "AVG." denotes average recognition rate; σ denotes standard deviation of the 10-cross-validation accuracies.
Berlin (a:fear, e:disgust, f:happiness, l:boredom, n:neutral, t:sadness, w:anger)

Test	Feature	Method	SN	Recognition Rate (%)							
				A	E	F	L	N	T	W	AVG. (σ)
#1	MS	RNN	No	62.50	30.00	32.60	66.80	57.20	77.10	77.70	61.04 (7.40)
	MFF			42.60	36.50	48.50	32.00	51.00	66.90	58.80	53.68 (9.44)
	ECC			52.50	43.00	42.60	32.50	48.40	54.60	66.30	51.73 (4.51)
	EFCC			44.00	51.50	39.80	34.70	46.60	61.20	74.90	51.60 (6.83)
	SMFCC			73.50	62.00	52.40	51.90	59.50	83.70	82.80	**68.12**(6.15)
#2	MS	SVM	No	61.95	71.22	43.66	86.32	70.38	83.28	86.08	69.38 (6.18)
	MFF			67.46	64.64	55.28	60.08	57.24	86.88	75.88	66.23 (5.85)
	ECC			68.90	28.45	65.30	67.60	56.43	71.23	82.34	64.18 (6.31)
	EFCC			75.36	64.72	56.46	59.63	58.15	67.35	85.84	63.78 (6.00)
	SMFCC			76.60	67.91	69.71	77.24	75.30	89.84	90.02	**76.28**(3.53)
#3	MS	RNN	Yes	53.30	39.00	40.90	73.00	72.10	85.30	67.20	64.04 (6.40)
	MFF			45.30	65.00	60.40	56.20	52.4	85.40	78.70	65.29 (8.57)
	ECC			59.50	43.50	55.30	43.10	50.00	70.60	74.60	58.37 (7.65)
	EFCC			53.80	52.50	46.10	43.30	51.60	69.50	72.40	57.11 (5.73)
	SMFCC			73.70	69.00	69.10	74.20	68.60	91.60	85.10	**76.81**(4.95)
#4	MS	SVM	Yes	67.30	55.71	48.06	81.53	81.88	86.88	88.49	72.41 (4.17)
	MFF			67.65	62.85	64.50	64.54	68.81	90.41	87.51	71.75 (3.91)
	ECC			66.43	55.55	67.01	61.93	60.68	68.68	78.55	62.48 (5.48)
	EFCC			68.07	52.36	51.82	68.53	61.12	77.58	79.05	61.91 (5.79)
	SMFCC			86.05	75.54	70.33	80.38	83.79	98.75	88.39	**81.13**(3.07)

The combination of different features extracted using EMD-TKEO method were performed in this group of comparative experiments. We present only the best combinations in Tables 4 and 5. We can see from Table 4 that the combination of SMFCC, MFF and MS features has better performance 85.63 % on the Berlin database. Furthermore, the combination of SMFCC, MFF, MS and EFCC is found to provide the best discrimination 90.72 % for the Spanish database (Table 5). It should also be noted that the Berlin database has limited content compared to the Spanish database, which limits the generalization of the obtained results.

Table 3: Recognition rates for different features on Spanish database;
Spanish (a:anger, d:disgust, f:fear, j:joy, n:neutral, s:surprise, t: sadness)

Test	Feature	Method	SN	Recognition Rate (%)							
				A	D	F	J	N	S	T	AVG. (σ)
#1	MS	RNN	No	59.60	60.90	56.30	51.70	72.00	57.90	85.30	64.71 (2.46)
	MFF			46.60	46.60	35.60	54.00	58.50	24.90	65.10	49.05 (3.49)
	ECC			55.60	65.40	49.20	65.20	74.50	46.90	77.20	63.93 (2.15)
	EFCC			57.40	62.30	42	61.10	75.00	42.00	75.40	61.74 (1.25)
	SMFCC			83.80	86.50	81.60	83.00	86.80	79.10	93.90	**87.80**(1.82)
#1	MS	SVM	No	72.21	76.18	71.98	61.88	83.78	64.82	86.67	75.13 (1.40)
	MFF			62.02	67.06	57.01	57.41	71.52	44.09	77.61	64.30 (2.73)
	ECC			64.78	72.56	60.28	68.59	79.61	52.01	84.89	70.92 (2.28)
	EFCC			64.53	71.56	58.91	64.25	78.98	50.38	82.98	68.79 (2.05)
	SMFCC			90.23	90.28	86.63	86.35	96.48	82.58	96.19	**90.94**(10.2)
#3	MS	RNN	Yes	61.20	66.00	64.30	52.40	79.80	53.80	77.50	68.25 (1.55)
	MFF			53.10	53.20	51.70	56.00	75.20	42.60	72.40	60.39 (2.52)
	ECC			53.80	63.80	52.80	60.60	76.80	46.60	82.90	64.67 (1.69)
	EFCC			51.10	60.80	54.10	59.60	74.50	42.50	82.50	62.73 (1.95)
	SMFCC			82.40	83.60	82.00	75.30	91.80	77.90	94.40	**85.09**(1.55)
#4	MS	SVM	Yes	67.85	75.27	72.31	61.17	81.14	60.74	85.52	72.91 (1.04)
	MFF			62.62	63.53	54.85	52.96	69.69	43.18	72.94	61.15 (1.54)
	ECC			61.63	68.57	59.72	63.41	76.48	48.06	80.78	66.82 (1.72)
	EFCC			59.84	68.63	55.77	60.10	76.14	46.82	79.64	65.43 (1.70)
	SMFCC			84.91	85.25	83.04	79.78	89.99	77.47	92.32	**85.71**(1.33)

Table 4: Recognition results using different combinations of features evaluated on the Berlin database

Method	Feature	SN	Average (avg)	Standard deviation (σ)
RNN	SMFCC+MFF+MS	No	73.81	7.28
	SMFCC+MFF+MS+EFCC		76.50	7.79
	all features		77.09	7.19
	SMFCC+MFF+MS	Yes	**83.14**	4.62
	SMFCC+MFF+MS+EFCC		82.34	4.91
	all features		82.97	4.19
SVM	SMFCC+MFF+MS	No	77.75	3.41
	SMFCC+MFF+MS+EFCC		80.21	4.54
	all features		80.41	4.84
	SMFCC+MFF+MS	Yes	**85.63**	4.30
	SMFCC+MFF+MS+EFCC		84.69	4.22
	all features		83.38	3.86

Table 5: Recognition results using different combinations of features evaluated on the Spanish database

Method	Feature	SN	Average (avg)	Standard deviation (σ)
RNN	SMFCC+MFF+MS	No	90.52	2.39
	SMFCC+MFF+MS+EFCC		**90.72**	1.79
	all features		90.65	1.85
	SMFCC+MFF+MS	Yes	88.18	0.47
	SMFCC+MFF+MS+EFCC		87.25	1.20
	all features		87.25	1.11
SVM	SMFCC+MFF+MS	No	89.98	1.76
	SMFCC+MFF+MS+EFCC		**90.69**	0.86
	all features		90.69	1.30
	SMFCC+MFF+MS	Yes	85.61	1.05
	SMFCC+MFF+MS+EFCC		85.67	1.69
	all features		85.77	1.02

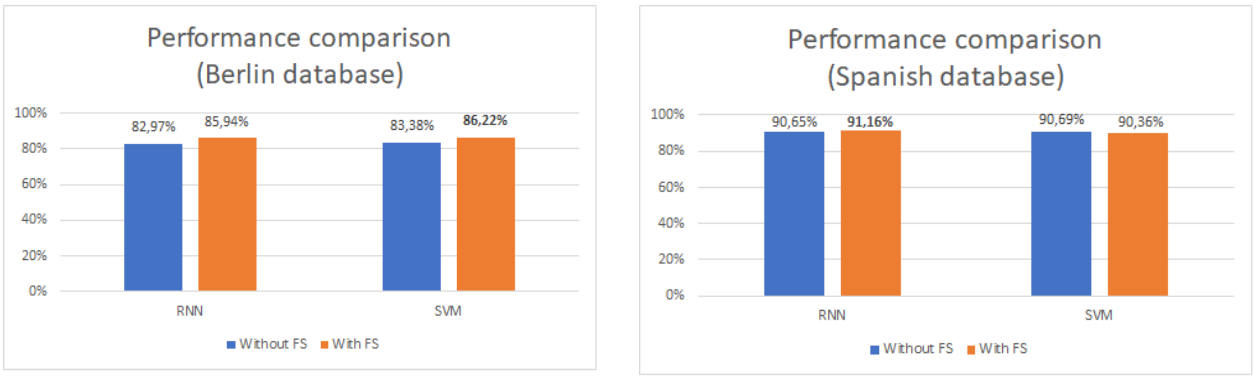
To investigate whether a smaller feature space leads to better recognition performance, we repeated all evaluations on the development set by applying a Recursive Feature Elimination (LR-RFE) for each modality combination. The stability of RFE depends heavily on the type of model that is used for feature ranking at each iteration. In our case, we tested the RFE based on an SVM and a regression models. We found that linear regression provides more stable results. The numeric results of applying LR-RFE feature selection technique are summarized in Table 6 with RNN and SVM employed for classification on both databases. We observed that SMFCC reported a best recognition rate of 83.20 %, which is about 5 percentages higher than the Fourier Transform based feature MFCC [33]. The combination of all features using LR-RFE feature selection gives the best results. The corresponding results of LR-RFE can be seen in Table 7. For the Spanish database, LR-RFE does not significantly improve the average accuracy. However, for recognition based on Berlin database, LR-RFE leads to a remarkable performance gain. This increase the average from 77.09 % to 85.94 % for RNN classifier. For the Spanish database, combination of all features after applying LR-RFE selection using RNN has the best recognition rate which is above 91.16%. The best accuracy achieved for Berlin database is 86.22 % using SVM classifier.

Table 6: Recognition rates for different features after applying LR-RFE feature selection method (Berlin and Spanish databases

Test	Feature	Method	SN	AVG. (σ)	
				Berlin	Spanish
#1	MS	RNN	No	60.82 (6.76)	64.55 (4.55)
	MFF			55.56 (9.41)	49.93 (3.27)
	ECC			53.70 (6.83)	62.87 (2.64)
	EFCC			53.58 (5.68)	60.50 (1.91)
	SMFCC			68.38 (5.92)	85.47 (2.17)
#2	MS	SVM	No	68.30 (6.33)	74.90 (1.79)
	MFF			65.09 (7.29)	64.86 (1.80)
	ECC			63.01 (7.12)	70.04 (1.31)
	EFCC			59.81 (7.28)	68.39 (1.93)
	SMFCC			78.11 (4.37)	**87.71 (1.45)**
#1	MS	RNN	Yes	66.27 (5.36)	66.11 (1.41)
	MFF			66.94 (8.69)	58.07 (1.88)
	ECC			62.01 (6.02)	64.47 (1.90)
	EFCC			60.33 (5.88)	61.78 (2.02)
	SMFCC			78.31 (3.36)	82.83 (1.50)
#1	MS	SVM	Yes	72.83 (4.63)	72.79 (1.48)
	MFF			72.45 (7.01)	62.16 (2.13)
	ECC			63.96 (4.30)	66.77 (1.37)
	EFCC			65.28 (6.84)	65.56 (1.39)
	SMFCC			**83.20 (5.51)**	83.34 (1.35)

Table 7: Recognition results with combination of all features using RNN and SVM before and after applying LR-RFE feature selection method (Berlin and Spanish databases)

SN	Classifier	LR-RFE	Berlin	Spanish
No	RNN	No	77.09 (7.19)	90.65 (1.85)
		Yes	85.94 (4.21)	**91.16** (1.79)
	SVM	No	80.41 (4.84)	90.69 (1.30)
		Yes	84.52 (5.61)	90.36 (1.05)
Yes	RNN	No	82.97 (4.19)	87.25 (1.11)
		Yes	85.94 (4.21)	88.39 (0.83)
	SVM	No	83.38 (3.86)	85.46 (1.02)
		Yes	**86.22** (4.27)	85.77 (1.13)



(a) (b)

Figure 16: Results for the SER system on both databases. Performance comparison between feature extraction with and without RFE Feature Selection (FS). Mean accuracy obtained using all features combined, for respectively the Berlin and Spanish databases, is shown

The performance comparison for two Machine Learning paradigms (RNN, SVM) without speaker normalization (SN), for the Berlin database, is shown in Figure 16a. For the Spanish database, the performance comparison using speaker normalization (SN) is shown in Figure 16b. The confusion matrix for the best recognition performance achieved by combination of all features on Spanish database, is shown in Table 8. The left-most column being the true emotions. The column indicates for a class, the number of correct predictions for this class and the number of samples confused with others class. The correct values are organized in a diagonal line from top-left to bottom-right of the matrix. For example, the total number of samples for emotion "Fear" in the data set is the sum of the values on the "Fear" column. Table 8 show that *anger* is most often confused with *joy* and the *neutral* emotion is most often confused with the *anger* emotional state.

Table 8: Confusion matrix for feature combination suing RNN classifier after LR-RFE selection based on Spanish database.

	Recognized emotion (%): 91.16 on average						
Emotion	Anger	Disgust	Fear	Joy	Neutral	Surprise	Sadness
Anger	**55**	1	1	11	1	3	0
Disgust	1	**68**	2	0	1	1	0
Fear	1	2	**64**	2	0	2	2
Joy	2	0	1	**67**	0	3	0
Neutral	3	1	1	1	**159**	0	0
Surprise	0	0	2	6	0	**64**	0
Sadness	0	1	0	0	0	0	**72**

8. Conclusion and future work

In this paper, we presented a speech emotion recognition (SER) system using two machine learning algorithms to classify seven emotions. In fact, we study how features impact recognition accuracy of emotions in speech. This work presents a novel features named modulation spectral (MS) features and modulation frequency features (MFF) based on AM-FM modulation model. These features are also combined with others. This work uses a nonlinear and non-stationary analysis based on the EMD technique. The combination of EMD and Teager-Kaiser Energy Operator (TKEO) show an important role for time-frequency analysis and particularly for AM-FM signal demodulation. Experimental study using Berlin and Spanish databases are carried out to evaluate the effectiveness of the proposed system. After feature extraction, we used a RFE method to select the best couple of features from the two databases and to improve the recognition rate. The conducted experiments clearly highlight the impact of the selected features in a classification approach. SER reported the best recognition rate of 91.16 % using RNN classifier on the Spanish database, and 86.22 % using SVM classifier on the Berlin database. From this result we can see that RNN often perform better with more data and it suffers from the problem of very long training times. Therefore, we concluded that the SVM model has a good potential for practical usage in comparison with RNN. In the future work, a speech emotion recognition system that combines multi-classifier will be investigated. We aim also to use other feature selection methods because the quality of the feature selection affects the emotion recognition rate: a good emotion feature selection method can select features reflecting emotion state quickly. We project also to improve the performance of our system by analyzing the correlation between set of features and emotions, because we have observed that for a specific emotion some features are more efficient than others. The overall aim of our work is to develop a system that will be used in a pedagogical interaction in classrooms, in order to help the teacher to orchestrate his class. For achieving this goal, we aim to test the system proposed in this work.

Appendix A. Notation

The Acronym	Signification
k	frame
N	number of IMF
N_f	number of frame
j	modulation band
$res_n(t)$	residue signal
$E_u(t)$	upper envelope
$E_l(t)$	lower envelope
$f_c\}$	frequency center
$x(t)$	input speech signal
$\mathbb{1}_\Omega(x)$	indicator function
F_n	instantaneous frequency
$r_i(t)$	the ith IMF of signal
A_n	instantaneous amplitude
L_F	length of frame in samples
$r(t)$	intrinsic mode function (IMF)
$h(f)$	marginal Hilbert spectrum (MHS)
$H(f,t)$	Hilbert spectrum (time-frequency distribution of the instantaneous energy envelope)

References

[1] A. Milton, S. S. Roy, S. T. Selvi, Svm scheme for speech emotion recognition using mfcc feature, International Journal of Computer Applications 69 (9).

[2] W. Zhang, D. Zhao, Z. Chai, L. T. Yang, X. Liu, F. Gong, S. Yang, Deep learning and svm-based emotion recognition from chinese speech for smart affective services, Software: Practice and Experience 47 (8) (2017) 1127–1138.

[3] R. Chen, Y. Zhou, Y. Qian, Emotion recognition using support vector machine and deep neural network, in: National Conference on Man-Machine Speech Communication, Springer, 2017, pp. 122–131.

[4] G. D. Sree, P. Chandrasekhar, B. Venkateshulu, Svm based speech emotion recognition compared with gmm-ubm and nn, International Journal of Engineering Science 3293.

[5] S. Prasomphan, Improvement of speech emotion recognition with neural network classifier by using speech spectrogram, in: Systems, Signals and Image Processing (IWSSIP), 2015 International Conference on, IEEE, 2015, pp. 73–76.

[6] W. Lim, D. Jang, T. Lee, Speech emotion recognition using convolutional and recurrent neural networks, in: Signal and Information Processing Association Annual Summit and Conference (APSIPA), 2016 Asia-Pacific, IEEE, 2016, pp. 1–4.

[7] S. Mirsamadi, E. Barsoum, C. Zhang, Automatic speech emotion recognition using recurrent neural networks with local attention, in: Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, IEEE, 2017, pp. 2227–2231.

[8] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. A. Mahjoub, A review on speech emotion recognition: Case of pedagogical interaction in classroom, in: Advanced Technologies for Signal and Image Processing (ATSIP), 2017 International Conference on, IEEE, 2017, pp. 1–7.

[9] L. Kerkeni, Y. Serrestou, M. Mbarki, M. Mahjoub, K. Raoof, C. Cleder, Speech emotion recognition: Recurrent neural networks compared to svm and linear regression, in: Artificial Neural Networks and Machine Learning, 2017. ICANN 2017. International Conference on, Springer, 2017, pp. 451–453.
[10] C.-C. Wang, Y. Kang, Feature extraction techniques of non-stationary signals for fault diagnosis in machinery systems, Journal of Signal and Information Processing 3 (01) (2012) 16.
[11] A. F. Haque, Frequency analysis and feature extraction of impressive tools, International Journal of Advance Innovations, Thoughts & Ideas 2 (2) (2013) 1.
[12] M. Nayak, B. S. Panigrahi, Advanced signal processing techniques for feature extraction in data mining, International Journal of Computer Applications 19 (9) (2011) 30–37.
[13] R. Fonseca-Pinto, A new tool for nonstationary and nonlinear signals: The hilbert-huang transform in biomedical applications, in: Biomedical Engineering, Trends in Electronics, Communications and Software, InTech, 2011.
[14] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, H. 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 A: Mathematical, Physical and Engineering Sciences 454 (1998) 903–995.
[15] X. Li, X. Li, Speech emotion recognition using novel hht-teo based features., JCP 6 (5) (2011) 989–998.
[16] X. Li, X. Li, X. Zheng, D. Zhang, Emd-teo based speech emotion recognition, Life System Modeling and Intelligent Computing (2010) 180–189.
[17] C. Shahnaz, S. Sultana, S. A. Fattah, R. M. Rafi, I. Ahmmed, W.-P. Zhu, M. O. Ahmad, Emotion recognition based on emd-wavelet analysis of speech signals, in: Digital Signal Processing (DSP), 2015 IEEE International Conference on, IEEE, 2015, pp. 307–310.
[18] N. Zhuang, Y. Zeng, L. Tong, C. Zhang, H. Zhang, B. Yan, Emotion recognition from eeg signals using multidimensional information in emd domain, BioMed research international 2017.
[19] R. Sharma, L. Vignolo, G. Schlotthauer, M. A. Colominas, H. L. Rufiner, S. Prasanna, Empirical mode decomposition for adaptive am-fm analysis of speech: A review, Speech Communication (2017).
[20] Z. Wu, N. E. Huang, Ensemble empirical mode decomposition: a noise-assisted data analysis method, Advances in adaptive data analysis 1 (01) (2009) 1–41.
[21] S. G. Goel, et al., Speech emotion recognition using eemd, svm & ann, Ph.D. thesis (2014).
[22] I. Antoniadou, G. Manson, N. Dervilis, T. Barszcz, W. Staszewski, K. Worden, Use of the teager-kaiser energy operator for condition monitoring of a wind turbine gearbox, in: International Conference on Noise and Vibration Engineering 2012, ISMA 2012, including USD 2012: International Conference on Uncertainty in Structure Dynamics, Vol. 6, 2012, pp. 4255–4268.
[23] K. Khaldi, A.-O. Boudraa, A. Komaty, Speech enhancement using empirical mode decomposition and the teager–kaiser energy operator, The Journal of the Acoustical Society of America 135 (1) (2014) 451–459.
[24] J. Tang, S. Alelyani, H. Liu, Feature selection for classification: A review, Data Classification: Algorithms and Applications (2014) 37.
[25] L. Kerkeni, Y. Serrestou, M. Mbarki, M. Mahjoub, K. Raoof, Speech emotion recognition: Methods and cases study, in: International Conference on Agents and Artificial Intelligence (ICAART), January 2018.
[26] P. Maragos, J. F. Kaiser, T. F. Quatieri, Energy separation in signal modulations with application to speech analysis, IEEE transactions on signal processing 41 (10) (1993) 3024–3051.
[27] A. Potamianos, P. Maragos, Speech analysis and synthesis using an am–fm modulation model1, Speech communication 28 (3) (1999) 195–209.
[28] V. Sethu, E. Ambikairajah, J. Epps, Empirical mode decomposition based weighted frequency feature for speech-based emotion classification, in: Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on, IEEE, 2008, pp. 5017–5020.
[29] N. E. Huang, Z. Shen, S. R. Long, M. C. Wu, H. H. Shih, Q. Zheng, N.-C. Yen, C. C. Tung, H. H. Liu, The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis, in: Proceedings of the Royal Society of London A: mathematical, physical and

engineering sciences, Vol. 454, The Royal Society, 1998, pp. 903–995.
[30] P. Maragos, J. F. Kaiser, T. F. Quatieri, On amplitude and frequency demodulation using energy operators, IEEE Transactions on signal processing 41 (4) (1993) 1532–1550.
[31] P. Maragos, T. F. Quatieri, J. F. Kaiser, Speech nonlinearities, modulations, and energy operators, in: Acoustics, Speech, and Signal Processing, 1991. ICASSP-91., 1991 International Conference on, IEEE, 1991, pp. 421–424.
[32] A. Potamianos, P. Maragos, A comparison of the energy operator and the hilbert transform approach to signal and speech demodulation, Signal processing 37 (1) (1994) 95–120.
[33] S. Wu, T. H. Falk, W.-Y. Chan, Automatic speech emotion recognition using modulation spectral features, Speech communication 53 (5) (2011) 768–785.
[34] L. Atlas, S. A. Shamma, Joint acoustic and modulation frequency, EURASIP Journal on Applied Signal Processing 2003 (2003) 668–675.
[35] D. W. Aha, R. L. Bankert, Feature selection for case-based classification of cloud types: An empirical comparison, in: Proceedings of the AAAI-94 workshop on Case-Based Reasoning, Vol. 106, 1994, p. 112.
[36] K.-B. Duan, J. C. Rajapakse, H. Wang, F. Azuaje, Multiple svm-rfe for gene selection in cancer classification with expression data, IEEE transactions on nanobioscience 4 (3) (2005) 228–234.
[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.
[38] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene selection for cancer classification using support vector machines, Machine learning 46 (1-3) (2002) 389–422.
[39] Y. Pan, P. Shen, L. Shen, Speech emotion recognition using support vector machine, International Journal of Smart Home 6 (2) (2012) 101–108.
[40] S. Wu, Recognition of human emotion in speech using modulation spectral features and support vector machines, Ph.D. thesis (2009).
[41] S. R. Gunn, et al., Support vector machines for classification and regression, ISIS technical report 14 (1) (1998) 5–16.
[42] Svm and kernel methods matlab toolbox, http://asi.insa-rouen.fr/enseignants/~arakoto/toolbox/.
[43] L. Zhang, S. Wang, B. Liu, Deep learning for sentiment analysis: A survey, arXiv preprint arXiv:1801.07883 (2018).
[44] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.
[45] S. Chen, Q. Jin, Multi-modal Dimensional Emotion Recognition using Recurrent Neural Networks, Brisbane, Australia, 2015.
[46]
[47] I. Saratxaga, E. Navas, I. Hernáez, I. Luengo, Designing and recording an emotional speech database for corpus based synthesis in basque, in: Proc. of fifth international conference on Language Resources and Evaluation (LREC), 2006, pp. 2126–2129.
[48] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, A Database of German Emotional Speech, INTERSPEECH (2005).
[49] Emotional speech synthesis database, http://catalog.elra.info/en-us/repository/browse/emotional-speech-synthesis-database/629db920a9e811e7a093ac9e1701ca021bdb22603cbc4702a3b6f592d250e427/.
[50] Z.-T. Liu, M. Wu, W.-H. Cao, J.-W. Mao, J.-P. Xu, G.-Z. Tan, Speech emotion recognition based on feature selection and extreme learning machine decision tree, Neurocomputing 273 (2018) 271–280.