Automatic Speech Emotion Recognition Using Machine Learning

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

Chapter

Automatic Speech Emotion Recognition Using Machine Learning

Leila Kerkeni, Youssef Serrestou, Mohamed Mbarki, Kosai Raoof, Mohamed Ali Mahjoub and Catherine Cleder

Abstract

This chapter presents a comparative study of speech emotion recognition (SER) systems. Theoretical definition, categorization of affective state and the modalities of emotion expression are presented. To achieve this study, an SER system, based on different classifiers and different methods for features extraction, is developed. Mel-frequency cepstrum coefficients (MFCC) and modulation spectral (MS) features are extracted from the speech signals and used to train different classifiers. Feature selection (FS) was applied in order to seek for the most relevant feature subset. Several machine learning paradigms were used for the emotion classification task. A recurrent neural network (RNN) classifier is used first to classify seven emotions. Their performances are compared later to multivariate linear regression (MLR) and support vector machines (SVM) techniques, which are widely used in the field of emotion recognition for spoken audio signals. Berlin and Spanish databases are used as the experimental data set. This study shows that for Berlin database all classifiers achieve an accuracy of 83% when a speaker normalization (SN) and a feature selection are applied to the features. For Spanish database, the best accuracy (94 %) is achieved by RNN classifier without SN and with FS.

Keywords: speech emotion recognition, feature extraction recurrent neural, network SVM, multivariate linear regression, MFCC, modulation spectral features, machine learning

1. Introduction

Emotion plays a significant role in daily interpersonal human interactions. This is essential to our rational as well as intelligent decisions. It helps us to match and understand the feelings of others by conveying our feelings and giving feedback to others. Research has revealed the powerful role that emotion play in shaping human social interaction. Emotional displays convey considerable information about the mental state of an individual. This has opened up a new research field called automatic emotion recognition, having basic goals to understand and retrieve desired emotions. In prior studies, several modalities have been explored to recognize the emotional states such as facial expressions [1], speech [2], physiological signals [3], etc. Several inherent advantages make speech signals a good source for affective computing. For example, compared to many other biological signals

(e.g., electrocardiogram), speech signals usually can be acquired more readily and economically. This is why the majority of researchers are interested in speech emotion recognition (SER). SER aims to recognize the underlying emotional state of a speaker from her voice. The area has received increasing research interest all through current years. There are many applications of detecting the emotion of the persons like in the interface with robots, audio surveillance, web-based E-learning, commercial applications, clinical studies, entertainment, banking, call centers, cardboard systems, computer games, etc. For classroom orchestration or E-learning, information about the emotional state of students can provide focus on the enhancement of teaching quality. For example, a 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 learner's emotional state should be considered in the classroom.

Three key issues need to be addressed for successful SER system, namely, (1) choice of a good emotional speech database, (2) extracting effective features, and (3) designing reliable classifiers using machine learning algorithms. In fact, the emotional feature extraction is a main issue in the SER system. Many researchers [4] have proposed important speech features which contain emotion information, such as energy, pitch, formant frequency, Linear Prediction Cepstrum Coefficients (LPCC), Mel-frequency cepstrum coefficients (MFCC), and modulation spectral features (MSFs) [5]. Thus, most researchers prefer to use combining feature set that is composed of many kinds of features containing more emotional information [6]. However, using a combining feature set may give rise to high dimension and redundancy of speech features; thereby, it makes the learning process complicated for most machine learning algorithms and increases the likelihood of overfitting. Therefore, feature selection is indispensable to reduce the dimensions redundancy of features. A review for feature selection models and techniques is presented in [7]. Both feature extraction and feature selection are capable of improving learning performance, lowering computational complexity, building better generalizable models, and decreasing required storage. The last step of speech emotion recognition is classification. It involves classifying the raw data in the form of utterance or frame of the utterance into a particular class of emotion on the basis of features extracted from the data. In recent years in speech emotion recognition, researchers proposed many classification algorithms, such as Gaussian mixture model (GMM) [8], hidden Markov model (HMM) [9], support vector machine (SVM) [10–14], neural networks (NN) [15], and recurrent neural networks (RNN) [16–18]. Some other types of classifiers are also proposed by some researchers such as a modified brain emotional learning model (BEL) [19] in which the adaptive neuro-fuzzy inference system (ANFIS) and multilayer perceptron (MLP) are merged for speech emotion recognition. Another proposed strategy is a multiple kernel Gaussian process (GP) classification [17], in which two similar notions in the learning algorithm are presented by combining the linear kernel and radial basis function (RBF) kernel. The Voiced Segment Selection (VSS) algorithm also proposed in [20] deals with the voiced signal segment as the texture image processing feature which is different from the traditional method. It uses the Log-Gabor filters to extract the voiced and unvoiced features from spectrogram to make the classification.

In previous work [21], we present a system for the recognition of «seven acted emotional states (anger, disgust, fear, joy, sadness, and surprise)». To do that, we extracted the MFCC and MS features and used them to train three different machine learning paradigms (MLR, SVM, and RNN). We demonstrated that the combination of both features has a high accuracy above 94% on the Spanish database. All previously published works generally use the Berlin database. To our

knowledge, the Spanish emotional database has never been used before. For this reason, we have chosen to compare them. In this chapter, we concentrate to improve accuracy; more experiments have been performed. This chapter mainly makes the following contributions:

- The effect of speaker normalization (SN) is also studied, which removes the mean of features and normalizes them to unit variance. Experiments are performed under a speaker-independent condition.

- Additionally, a feature selection technique is assessed to obtain good features from the set of features extracted in [21].

The rest of the chapter is organized as follows. In the next section, we start by introducing the nature of speech emotions. Section 3 describes features we extracted from a speech signal. A feature selection method and machine learning algorithms used for SER are presented. Section 4 reports on the databases we used and presents the simulation results obtained using different features and different machine learning (ML) paradigms. Section 5 closes this chapter by analyses and conclusion.

2. Emotion and classification

This section is concerned with defining the term emotion, presenting its different models. Also for recognizing emotions, there are several techniques and inputs that can be used. A brief description of all of the techniques is presented here.

2.1 Definition

A definition is both important and difficult because the everyday word "emotion" is a notoriously fluid term in meaning. Emotion is one of the most difficult concepts to define in psychology. In fact, there are different definitions of emotions in the scientific literature. In everyday speech, emotion is any relatively brief conscious experience characterized by intense mental activity and a high degree of pleasure or displeasure [22, 23]. Scientific discourse has drifted to other meanings and there is no consensus on a definition. Emotion is often entwined with temperament, mood, personality, motivation, and disposition. In psychology, emotion is frequently defined as a complex state of feeling that results in physical and psychological changes. These changes influence thought and behavior. According to other theories, emotions are not causal forces but simply syndromes of components such as motivation, feeling, behavior, and physiological changes [24]. In 1884, in *What is an emotion?* [25], American psychologist and philosopher William James proposed a theory of emotion whose influence was considerable. According to his thesis, the feeling of intense emotion corresponds to the perception of specific bodily changes. This approach is found in many current theories: the bodily reaction is the cause and not the consequence of the emotion. The scope of this theory is measured by the many debates it provokes. This illustrates the difficulty of agreeing on a definition of this dynamic and complex phenomenon that we call emotion. "Emotion" refers to a wide range of affective processes such as moods, feelings, affects, and well-being [26]. The term "emotion" in [6] has been also referred to an extremely complex state associated with a wide variety of mental, physiological, and physical events.

2.2 Categorization of emotions

The categorization of emotions has long been a hot subject of debate in different fields of psychology, affective science, and emotion research. It is mainly based on two popular approaches: categorical (termed discrete) and dimensional (termed continuous). In the first approach, emotions are described with a discrete number of classes. Many theorists have conducted studies to determine which emotions are basic [27]. A most popular example is Ekman [28] who proposed a list of six basic emotions, which are anger, disgust, fear, happiness, sadness, and surprise. He explains that each emotion acts as a discrete category rather than an individual emotional state. In the second approach, emotions are a combination of several psychological dimensions and identified by axes. Other researchers define emotions according to one or more dimensions. Wilhelm Max Wundt proposed in 1897 that emotions can be described by three dimensions: (1) strain versus relaxation, (2) pleasurable versus unpleasurable, and (3) arousing versus subduing [29]. PAD emotional state model is another three-dimensional approach by Albert Mehrabian and James Russell where PAD stands for pleasure, arousal, and dominance. Another popular dimensional model was proposed by James Russell in 1977. Unlike the earlier three-dimensional models, Russell's model features only two dimensions which include (1) arousal (or activation) and (2) valence (or evaluation) [29].

The categorical approach is commonly used in SER [30]. It characterizes emotions used in everyday emotion words such as joy and anger. In this work, a set of six basic emotions (anger, disgust, fear, joy, sadness, and surprise) plus neutral, corresponding to the six emotions of Ekman's model, were used for the recognition of emotion from speech using the categorical approach.

2.3 Sensory modalities for emotion expression

There is vigorous debate about what exactly individual can express nonverbally. Humans can express their emotions through many different types of nonverbal communication including facial expressions, quality of speech produced, and physiological signals of the human body. In this section, we discuss each of these categories.

2.3.1 Facial expressions

The human face is extremely expressive, able to express countless emotions without saying a word [31]. And unlike some forms of nonverbal communication, facial expressions are universal. The facial expressions for happiness, sadness, anger, surprise, fear, and disgust are the same across cultures.

2.3.2 Speech

In addition to faces, voices are an important modality for emotional expression. Speech is a relevant communicational channel enriched with emotions: the voice in speech not only conveys a semantic message but also the information about the emotional state of the speaker. Some important voice feature vectors that have been chosen for research such as fundamental frequency, mel-frequency cepstral coefficient (MFCC), prediction cepstral coefficient (LPCC), etc.

2.3.3 Physiological signals

The physiological signals related to autonomic nervous system allow to assess objectively emotions. These include electroencephalogram (EEG), heart rate (HR),

electrocardiogram (ECG), respiration (RSP), blood pressure (BP), electromyogram (EMG), skin conductance (SC), blood volume pulse (BVP), and skin temperature (ST) [32]. Using physiological signals to recognize emotions is also helpful to those people who suffer from physical or mental illness thus exhibit problems with facial expressions or tone of voice.

3. Speech emotion recognition (SER) system

3.1 Block diagram

Our SER system consists of four main steps. First is the voice sample collection. The second features vector that is formed by extracting the features. As the next step, we tried to determine which features are most relevant to differentiate each emotion. These features are introduced to machine learning classifier for recognition. This process is described in **Figure 1**.

3.2. Feature extraction

The speech signal contains a large number of parameters that reflect the emotional characteristics. One of the sticking points in emotion recognition is what features should be used. In recent research, many common features are extracted, such as energy, pitch, formant, and some spectrum features such as linear prediction coefficients (LPC), mel-frequency cepstrum coefficients (MFCC), and modulation spectral features. In this work, we have selected modulation spectral features and MFCC, to extract the emotional features.

Mel-frequency cepstrum coefficient (MFCC) is the most used representation of the spectral property of voice signals. These are the best for speech recognition as it takes human perception sensitivity with respect to frequencies into consideration. For each frame, the Fourier transform and the energy spectrum were estimated and mapped into the Mel-frequency scale. The discrete cosine transform (DCT) of the Mel log energies was estimated, and the first 12 DCT coefficients provided the

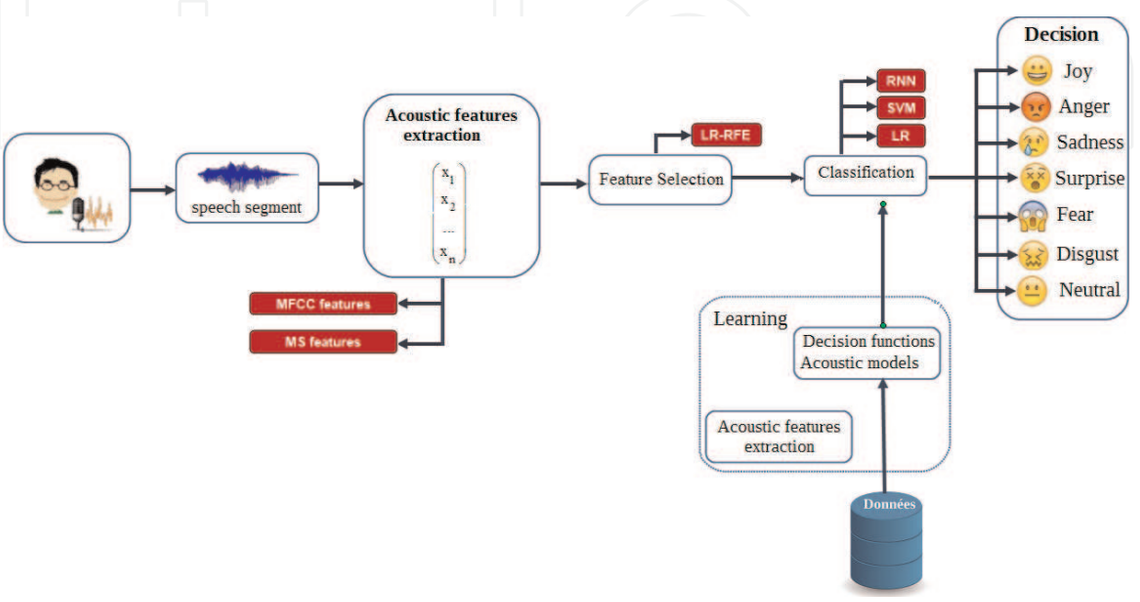

Figure 1.
Block diagram of the proposed system.

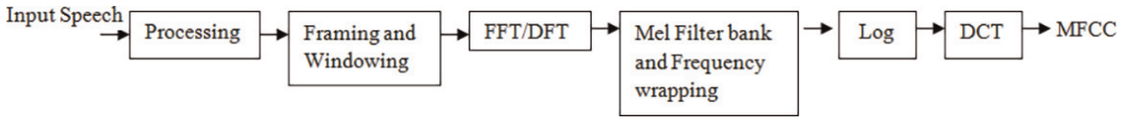

Figure 2.
Schema of MFCC extraction [33].

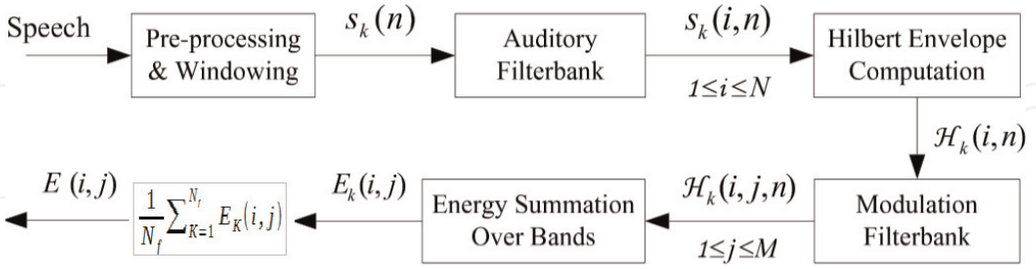

Figure 3.
Process for computing the ST representation [5].

MFCC values used in the classification process. Usually, the process of calculating MFCC is shown in **Figure 2**.

In our research, we extract the first 12 order of the MFCC coefficients where the speech signals are sampled at 16 KHz. For each order coefficients, we calculate the mean, variance, standard deviation, kurtosis, and skewness, and this is for the other all the frames of an utterance. Each MFCC feature vector is 60-dimensional.

Modulation spectral features (MSFs) are extracted from an auditory-inspired long-term spectro-temporal representation. These features are obtained by emulating the spectro-temporal (ST) processing performed in the human auditory system and consider regular acoustic frequency jointly with modulation frequency. The steps for computing the ST representation are illustrated in **Figure 3**. In order to obtain the ST representation, the speech signal is first decomposed by an auditory filterbank (19 filters in total). The Hilbert envelopes of the critical-band outputs are computed to form the modulation signals. A modulation filterbank is further applied to the Hilbert envelopes to perform frequency analysis. The spectral contents of the modulation signals are referred to as modulation spectra, and the proposed features are thereby named modulation spectral features (MSFs) [5]. Lastly, the ST representation is formed by measuring the energy of the decomposed envelope signals, as a function of regular acoustic frequency and modulation frequency. The energy, taken over all frames in every spectral band, provides a feature. In our experiment, an auditory filterbank with $N = 19$ filters and a modulation filterbank with $M = 5$ filters are used. In total, 95 (19×5) MSFs are calculated in this work from the ST representation.

3.3 Feature selection

As reported by Aha and Bankert [34], the objective of feature selection in ML is to "reduce the number of features used to characterize a dataset so as to improve a learning algorithm's performance on a given task." The objective will be the maximization of the classification accuracy in a specific task for a certain learning algorithm; as a collateral effect, the number of features to induce the final classification model will be reduced. Feature selection (FS) aims to choose a subset of the relevant features from the original ones according to certain relevance evaluation criterion, which usually leads to higher recognition accuracy [35]. It can drastically reduce the running time of the learning algorithms. In this section, we present an

effective feature selection method used in our work, named recursive feature elimination with linear regression (LR-RFE).

Recursive feature elimination (RFE) uses a model (e.g., 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 the current set of features. That 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]. Other research also uses RFE with another linear model such as SVM-RFE that is an SVM-based feature selection algorithm created by [38]. Using SVM-RFE, Guyon et al. selected key and important feature sets. In addition to improving the classification accuracy rate, it can reduce classification computational time.

3.4 Classification methods

Many machine learning algorithms have been used for discrete emotion classification. The goal of these algorithms is to learn from the training samples and then use this learning to classify new observation. In fact, there is no definitive answer to the choice of the learning algorithm; every technique has its own advantages and limitations. For this reason, here we chose to compare the performance of three different classifiers.

Multivariate linear regression classification (MLR) is a simple and efficient computation of machine learning algorithms, and it can be used for both regression and classification problems. We have slightly modified the LRC algorithm described as follow Algorithm 1 [39]. We calculated (in step 3) the absolute value of the difference between original and predicted response vectors ($|y - y_i|$), instead of the Euclidean distance between them ($\|y - y_i\|$).

Support vector machines (SVM) are an optimal margin classifier in machine learning. It is also used extensively in many studies that related to audio emotion recognition which can be found in [10, 13, 14]. It can have a very good classification performance compared to other classifiers especially for limited training data [11]. SVM theoretical background can be found in [40]. A MATLAB toolbox implementing SVM is freely available in [41]. A polynomial kernel is investigated in this work.

Algorithm 1. Linear Regression Classification (LRC)

Inputs: Class models $X_i \in \mathbb{R}^{q \times p_i}$, $i = 1, 2, ..., N$ and a test speech vector $y \in \mathbb{R}^{q \times 1}$
Output: Class of y

1. $\hat{\beta}_i \in \mathbb{R}^{p_i \times 1}$ is evaluated against each class model, $\hat{\beta}_i = \left(X_i^T X_i\right)^{(-1)} X_i^T y$, $i = 1, 2, ..., N$
2. $\hat{y}_i$ is computed for each $\hat{\beta}_i$, $\hat{y}_i = X_i \hat{\beta}_i, \quad i = 1, 2, ..., N$;
3. Distance calculation between original and predicted response variables $d_i(y) = |y - y_i|, \quad i = 1, 2, ..., N$;
4. Decision is made in favor of the class with the minimum distance $d_i(y)$

Recurrent neural networks (RNN) are suitable for learning time series data, and it has shown improved performance for classification task [42]. While RNN

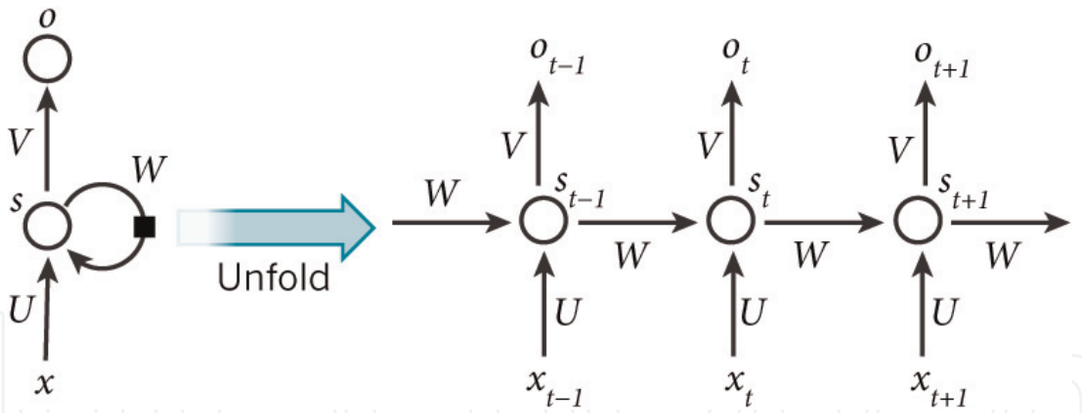

Figure 4.
A basic concept of RNN and unfolding in time of the computation involved in its forward computation [18].

models are effective at learning temporal correlations, they suffer from the vanishing gradient problem which increases with the length of the training sequences. To resolve this problem, long short-term memory (LSTM) RNNs were proposed by Hochreiter et al. [43]; it uses memory cells to store information so that it can exploit long-range dependencies in the data [17].

Figure 4 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 are presented in **Figure 4**) across all steps. The hidden state formulas and variables are as follows:

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

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

4. Experimental results and analysis

4.1 Emotional speech databases

The performance and robustness of the recognition systems will be easily affected if it is not well trained with a suitable database. Therefore, it is essential to have sufficient and suitable phrases in the database to train the emotion recognition system and subsequently evaluate its performance. There are three main types of databases: acted emotions, natural spontaneous emotions, and elicited emotions [27, 44]. In this work, we used an acted emotion databases because they contain strong emotional expressions. The literature on speech emotion recognition [45] shows that the majority of studies have been conducted with emotional acted speech. In this section, we detailed the two emotional speech databases used for classifying discrete emotions in our experiments: Berlin Database and Spanish Database.

4.2 Berlin database

The Berlin database [46] is widely used in emotional speech recognition. It contains 535 utterances spoken by 10 actors (5 female, 5 male) in 7 simulated emotions (anger, boredom, disgust, fear, joy, sadness, and neutral). This database was chosen for the following reasons: (i) the quality of its recording is very good, and (ii) it is public [47] and popular database of emotion recognition that is recommended in the literature [19].

4.3 Spanish database

The INTER1SP Spanish emotional database contains utterances from two professional actors (one female and one male speaker).The Spanish corpus that we have the right to access (free for academic and research use) [48] was recorded twice in the «six basic emotions plus neutral (anger, sadness, joy, fear, disgust, surprise and neutral/normal)». Four additional neutral variations (soft, loud, slow, and fast) were recorded once. This is preferred to other created database because it is available for researchers use and it contains more data (6041 utterances in total). This paper has focused on only seven main emotions from the Spanish database in order to achieve a higher and more accurate rate of recognition and to make the comparison with the Berlin database detailed above.

4.4 Results and analysis

In this section, experimentation results are presented and discussed. We report the recognition accuracy of using MLR, SVM, and RNN classifiers. Experimental evaluation is performed on the Berlin and Spanish databases. All classification results are obtained under tenfold 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 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. 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 and in some cases prevent 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 [6]. SN is useful to compensate for the variations due to speaker diversity rather than the change of emotional state. We used in this section the SN scheme that has given the best results in [6]. The features of each speaker are normalized with a mean of 0 and a standard deviation of 1. **Tables 1–3** show the recognition rate for each combination of various features and classifiers based on Berlin and Spanish databases. These experiments use feature set without feature selection. As shown in **Table 1**, SVM classifier yields better results above 81%, with feature combination of MFCC and MS for Berlin database. Our results have improved compared to previous results in [21] because we changed the SVM parameters for each type of features to develop a good model.

From **Table 1**, it can be concluded that applying SN improves recognition results for Berlin database. But this is not the case for the Spanish database, as demonstrated in **Tables 2** and **3**. Results are the same with the three different classifiers. This can be explained by the number of speakers in each database. The Berlin database contains 10 different speakers, compared to the Spanish database that contains only two speakers and probably the language impact. As regarding the RNN method, we found that combining both types of features has the worst recognition rate for the Berlin database, as shown in **Table 3**. That is because the RNN model has too many parameters (155 coefficients in total) and a poor training data. This is the phenomena of overfitting. This is confirmed by the fact that when we reduced the number of features from 155 to 59 features, the results show an increase of above 13%, as shown in **Table 4**. 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

				Recognition rate (%)								
Test	**Feature**	**Method**	**SN**	**A**	**E**	**F**	**L**	**N**	**T**	**W**	**AVG.**	**(σ)**
#1	MS	MLR	No	45.90	45.72	48.78	77.08	59.43	79.91	75.94	66.23	(5.85)
	MFCC			56.55	62.28	45.60	54.97	57.35	74.36	91.37	64.70	(3.20)
	MFCC+SM			70.26	73.04	51.95	82.44	69.55	82.49	76.55	73.00	(3.23)
#2	MS	SVM	No	56.61	54.78	51.17	70.98	67.32	67.50	73.13	70.63	(6.45)
	MFCC			73.99	64.14	64.76	55.30	62.28	84.13	83.13	71.70	(4.24)
	MFCC+SM			82.03	68.70	69.09	79.16	76.99	80.89	80.63	81.10	(2.73)
#3	MS	MLR	Yes	48.98	35.54	32.66	80.35	55.54	88.79	85.77	64.20	(5.27)
	MFCC			59.71	59.72	48.65	67.10	67.98	91.73	87.51	71.00	(4.19)
	MFCC+SM			72.32	68.82	51.98	82.60	81.72	91.96	80.71	75.25	(2.49)
#4	MS	SVM	Yes	62.72	49.44	37.29	76.14	71.30	88.44	80.15	71.90	(2.38)
	MFCC			70.68	56.55	56.99	59.88	68.14	91.88	85.44	77.60	(4.35)
	MFCC+SM			77.37	69.67	58.16	79.87	88.57	98.75	86.64	81.00	(2.45)

Berlin (a, fear; e, disgust; f, happiness; l, boredom; n, neutral; t, sadness; w, anger).

Table 1.
Recognition results with MS, MFCC features, and their combination on Berlin database; AVG. denotes average recognition rate; σ denotes standard deviation of the 10-cross-validation accuracies.

				Recognition rate (%)								
Test	**Feature**	**Method**	**SN**	**A**	**D**	**F**	**J**	**N**	**S**	**T**	**AVG.**	**(σ)**
#1	MS	MLR	No	67.72	44.04	68.78	46.95	89.58	63.10	78.49	69.22	(1.37)
	MFCC			67.85	61.41	75.97	60.17	95.79	71.89	84.94	77.21	(0.76)
	MFCC+SM			78.75	78.18	80.68	63.84	96.80	82.44	89.01	83.55	(0.55)
#2	MS	SVM	No	70.33	69.38	78.09	60.97	89.25	69.38	85.95	80.98	(1.09)
	MFCC			79.93	79.02	81.81	75.71	93.77	80.15	92.01	90.94	(0.93)
	MFCC+SM			84.90	88.26	89.44	80.90	96.58	83.89	95.63	89.69	(0.62)
#3	MS	MLR	Yes	64.76	49.02	66.87	44.52	87.50	58.26	78.70	67.84	(1.27)
	MFCC			66.54	57.83	74.56	56.98	94.02	72.32	89.63	76.47	(1.51)
	MFCC+SM			77.01	78.45	80.50	64.18	94.42	80.14	91.29	83.03	(0.97)
#4	MS	SVM	Yes	69.81	70.35	75.44	52.60	86.77	66.94	82.57	78.40	(1.64)
	MFCC			77.45	77.41	80.99	69.47	91.89	75.17	93.50	87.47	(0.95)
	MFCC+SM			85.28	84.54	84.49	73.47	93.43	81.79	94.04	86.57	(0.72)

Spanish (a, anger; d, disgust; f, fear; j, joy; n, neutral; s, surprise; t, sadness).

Table 2.
Recognition results with MS, MFCC features, and their combination on Spanish database.

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 regression models; we found that using linear regression provides more stable results. We observed from the previous results that the combination of the features gives the best results. So we applied LR-RFE feature selection only for this combination to improve accuracy. In this work, a total of 155 features were used;

Dataset	Feature	SN	Average (avg)	Standard deviation (σ)
Berlin	MS	No	66.32	5.93
	MFCC		69.55	3.91
	MFCC+MS	Yes	63.67	7.74
	MS		68.94	5.65
	MFCC		73.08	5.17
	MFCC+MS		76.98	4.79
Spanish	MS	No	82.30	2.88
	MFCC		86.56	2.80
	MFCC+MS		90.05	1.64
	MS	Yes	82.14	1.67
	MFCC		86.21	1.22
	MFCC+MS		87.02	0.36

Table 3.
Recognition results using RNN classifier based on Berlin and Spanish databases.

best features were chosen from feature selection. Fifty-nine features were selected by RFE feature selection method based on LR from the Berlin database and 110 features from the Spanish database. The corresponding results of LR-RFE can be seen in **Table 4**. For most setting using the Spanish database, LR-RFE does not significantly improve the average accuracy. However, for recognition based on Berlin database using the three classifiers, LR-RFE leads to a remarkable performance gain, as shown in **Figure 5**. This increases the average of MFCC combined with MS features from 63.67 to 78.11% for RNN classifier. These results are illustrated in **Table 4**. For the Spanish database, the feature combination of MFCC and MS after applying LR-RFE selection using RNN has the best recognition rate which is above 94.01%.

SN	Classifier	LR-RFE	Berlin	Spanish
No	MLR	No	73.00 (3.23)	83.55 (0.55)
		Yes	79.40 (3.09)	84.19 (0.96)
	SVM	No	81.10 (2.73)	89.69 (0.62)
		Yes	80.90 (3.17)	90.05 (0.80)
	RNN	No	63.67 (7.74)	90.05 (1.64)
		Yes	78.11 (3.53)	**94.01** (0.76)
Yes	MLR	No	75.25 (2.49)	83.03 (0.97)
		Yes	**83.20** (3.25)	82.27 (1.12)
	SVM	No	81.00 (2.45)	86.57 (0.72)
		Yes	**83.90** (2.46)	86.47 (1.34)
	RNN	No	76.98 (4.79)	87.02 (0.36)
		Yes	**83.42** (0.70)	85.00 (0.93)

Table 4.
Recognition results with combination of MFCC and MS features using ML paradigm before and after applying LR-RFE feature selection method (Berlin and Spanish databases).

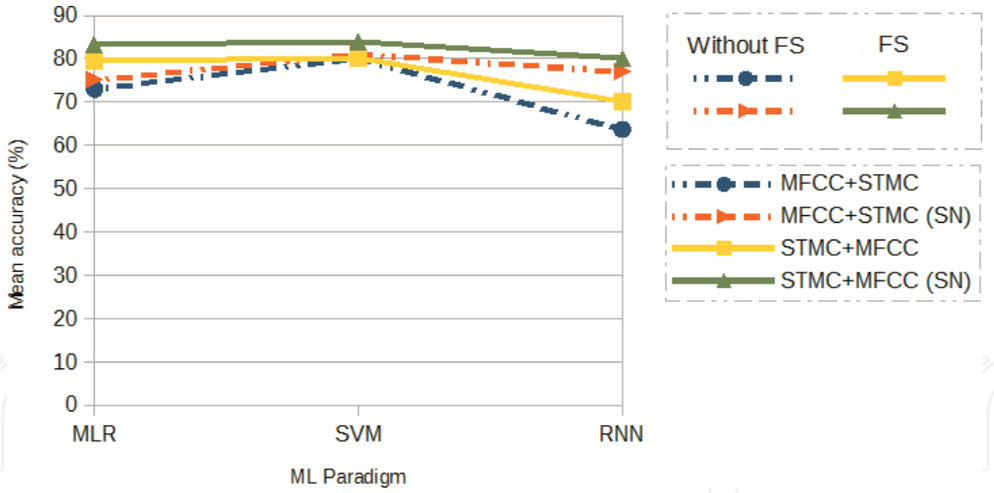

Figure 5.
Performance comparison of three machine learning paradigms (MLR, SVM, RNN) using speaker normalization (SN) and RFE feature selection (FS), for the Berlin database, is shown.

Emotion	Anger	Disgust	Fear	Joy	Neutral	Surprise	Sadness	Rate (%)
Anger	**79**	**1**	**0**	**1**	**2**	**3**	**0**	**91.86**
Disgust	0	**67**	**3**	**0**	**1**	**0**	**1**	**93.05**
Fear	0	3	**70**	**0**	**1**	**0**	**2**	**93.33**
Joy	3	1	1	**71**	**0**	**0**	**0**	**93.42**
Neutral	2	0	1	0	**156**	**0**	**1**	**97.50**
surprise	2	1	0	3	0	**60**	**0**	**92.30**
Sadness	0	0	1	0	2	0	**66**	**95.65**
Precision (%)	91.86	91.78	92.10	94.66	96.29	95.23	94.28	

Table 5.
Confusion matrix for feature combination after LR-RFE selection based on Spanish database.

The confusion matrix for the best recognition of emotions using MFCC and MS features with RNN based on Spanish database is shown in **Table 5**. The rate column lists per class recognition rates and precision for a class are the number of samples correctly classified divided by the total number of samples classified to the class. It can be seen that *Neutral* was the emotion that was least difficult to recognize from speech as opposed to *Disgust* which was the most difficult and it forms the most notable confusion pair with *Fear*.

5. Conclusion

In this current study, we presented an automatic speech emotion recognition (SER) system using three machine learning algorithms (MLR, SVM, and RNN) to classify seven emotions. Thus, two types of features (MFCC and MS) were extracted from two different acted databases (Berlin and Spanish databases), and a combination of these features was presented. In fact, we study how classifiers and features impact recognition accuracy of emotions in speech. A subset of highly discriminant features is selected. Feature selection techniques show that more information is not always good in machine learning applications. The machine learning models were trained and evaluated to recognize emotional states from

these features. SER reported the best recognition rate of 94% on the Spanish database using RNN classifier without speaker normalization (SN) and with feature selection (FS). For Berlin database, all of the classifiers achieve an accuracy of 83% when a speaker normalization (SN) and a feature selection (FS) are applied to the features. 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 and MLR models have a good potential for practical usage for limited data in comparison with RNN .

Enhancement of the robustness of emotion recognition system is still possible by combining databases and by fusion of classifiers. The effect of training multiple emotion detectors can be investigated by fusing these into a single detection system. 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. 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.

Author details

Leila Kerkeni[1,2*], Youssef Serrestou[1], Mohamed Mbarki[3], Kosai Raoof[1], Mohamed Ali Mahjoub[2] and Catherine Cleder[4]

1 LAUM UMR CNRS 6613, Le Mans Université, France

2 LATIS Lab, ENISo Université de Sousse, Tunisia

3 ISSAT, Université de Sousse, Tunisia

4 CREN Lab, Université de Nantes, France

*Address all correspondence to: kerkeni.leila@gmail.com

References

[1] Ali H, Hariharan M, Yaacob S, Adom AH. Facial emotion recognition using empirical mode decomposition. Expert Systems with Applications. 2015;**42**(3): 1261-1277

[2] Liu ZT, Wu M, Cao WH, Mao JW, Xu JP, Tan GZ. Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing. 2018;**273**: 271-280

[3] Ragot M, Martin N, Em S, Pallamin N, Diverrez JM. Emotion recognition using physiological signals: Laboratory vs. wearable sensors. In: International Conference on Applied Human Factors and Ergonomics. Springer; 2017. pp. 15-22

[4] Surabhi V, Saurabh M. Speech emotion recognition: A review. International Research Journal of Engineering and Technology (IRJET). 2016;**03**:313-316

[5] Wu S, Falk TH, Chan WY. Automatic speech emotion recognition using modulation spectral features. Speech Communication. 2011;**53**:768-785

[6] Wu S. Recognition of human emotion in speech using modulation spectral features and support vector machines [PhD thesis]. 2009

[7] Tang J, Alelyani S, Liu H. Feature selection for classification: A review. Data Classification: Algorithms and Applications. 2014:37

[8] Martin V, Robert V. Recognition of emotions in German speech using Gaussian mixture models. LNAI. 2009; **5398**:256-263

[9] Ingale AB, Chaudhari D. Speech emotion recognition using hidden Markov model and support vector machine. International Journal of Advanced Engineering Research and Studies. 2012:316-318

[10] Milton A, Sharmy Roy S, Tamil Selvi S. SVM scheme for speech emotion recognition using MFCC feature. International Journal of Computer Applications. 2013;**69**

[11] Divya Sree GS, Chandrasekhar P, Venkateshulu B. SVM based speech emotion recognition compared with GMM-UBM and NN. IJESC. 2016;**6**

[12] Melki G, Kecman V, Ventura S, Cano A. OLLAWV: Online learning algorithm using worst-violators. Applied Soft Computing. 2018;**66**:384-393

[13] Pan Y, Shen P, Shen L. Speech emotion recognition using support vector machine. International Journal of Smart Home. 2012;**6**:101-108

[14] Peipei S, Zhou C, Xiong C. Automatic speech emotion recognition using support vector machine. IEEE. 2011;**2**:621-625

[15] Sathit P. Improvement of speech emotion recognition with neural network classifier by using speech spectrogram. International Conference on Systems, Signals and Image Processing (IWSSIP). 2015:73-76

[16] Alex G, Navdeep J. Towards end-to-end speech recognition with recurrent neural networks. In: International Conference on Machine Learning. Vol. 32. 2014

[17] Chen S, Jin Q. Multi-Modal Dimensional Emotion Recognition using Recurrent Neural Networks. Australia: Brisbane; 2015

[18] Lim W, Jang D, Lee T. Speech emotion recognition using convolutional and recurrent neural networks. Asia-Pacific. 2017:1-4

[19] Sara M, Saeed S, Rabiee A. Speech Emotion Recognition Based on a Modified Brain Emotional Learning Model. Biologically inspired cognitive architectures. Elsevier; 2017;**19**:32-38

[20] Yu G, Eric P, Hai-Xiang L, van den HJ. Speech emotion recognition using voiced segment selection algorithm. ECAI. 2016;**285**:1682-1683

[21] Kerkeni L, Serrestou Y, Mbarki M, Mahjoub M, Raoof K. Speech emotion recognition: Methods and cases study. In: International Conference on Agents and Artificial Intelligence (ICAART); 2018

[22] Cabanac M. What is emotion? Behavioural Processes. 2002;**60**(2):69-83

[23] Schacter DL, Gilbert DT, Wegner DM. Psychology (2nd Edition). New York: Worth; 2011

[24] Barrett LF, Russell JA. The Psychological Construction of Emotion. Guilford Publications; 2014

[25] James W. What is an emotion? Mind. 1884;**9**(34):188-205

[26] Boekaerts M. The Crucial Role of Motivation and Emotion in Classroom Learning. The Nature of Learning: Using Research to Inspire Practice 2010. Paris: OECD Publishing; pp. 91-111

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

[28] Ekman P. An argument for basic emotions. Cognition & Emotion. 1992;**6**(3–4):169-200

[29] Matilda S. Emotion recognition: A survey. International Journal of Advanced Computer Research. 2015;**3**(1):14-19

[30] Koolagudi SG, Rao KS. Emotion recognition from speech: A review. International Journal of Speech Technology. 2012;**15**(2):99-117

[31] Schirmer A, Adolphs R. Emotion perception from face, voice, and touch: Comparisons and convergence. Trends in Cognitive Sciences. 2017;**21**(3): 216-228

[32] He C, Yao Yj, Ye Xs. An emotion recognition system based on physiological signals obtained by wearable sensors. In: Wearable Sensors and Robots. Springer; 2017. pp. 15-25

[33] Srinivasan V, Ramalingam V, Arulmozhi P. Artificial Neural Network Based Pathological Voice Classification Using MFCC Features. International Journal of Science, Environment and Technology (Citeseer). 2014;**3**:291-302

[34] Aha DW, Bankert RL. 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

[35] Song P, Zheng W. Feature selection based transfer subspace learning for speech emotion recognition. IEEE Transactions on Affective Computing. 2018

[36] Duan KB, Rajapakse JC, Wang H, Azuaje F. Multiple SVM-RFE for gene selection in cancer classification with expression data. IEEE Transactions on NanoBioscience. 2005;**4**(3):228-234

[37] Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. SCIKIT-learn: Machine learning in Python. Journal of Machine Learning Research. 2011;**12**:2825-2830

[38] Guyon I, Weston J, Barnhill S, Vapnik V. Gene selection for cancer classification using support vector

machines. Machine Learning. 2002;**46** (1–3):389-422

[39] Naseem I, Togneri R, Bennamoun M. Linear regression for face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2010;**32**:2106-2112

[40] Gunn SR. Support vector machines for classification and regression [PhD thesis]. 1998

[41] SVM and Kernel Methods MATLAB Toolbox. Available from: http://asi.insa-rouen.fr/enseignants/~arakoto/toolbox/

[42] Parthasarathy S, Tashev I. Convolutional neural network techniques for speech emotion recognition. In: 2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC). IEEE; 2018. pp. 121-125

[43] Sepp H, Jurgen S. Long Short-term Memory. Neural Computation. 1997;**9**: 1735-1780

[44] Vaudable C. Analyse et reconnaissance des émotions lors de conversations de centres d'appels [PhD thesis]. Université Paris Sud-Paris XI; 2012

[45] Swain M, Routray A, Kabisatpathy P. Databases, features and classifiers for speech emotion recognition: A review. International Journal of Speech Technology. 2018;**21**:1-28

[46] Burkhardt F, Paeschke A, Rolfes M, Sendlmeier W, Weiss B. A Database of German Emotional Speech. INTERSPEECH; 2005

[47] Berlin Database of Emotional Speech. Available from: http://emodb.bilderbar.info/start.html

[48] Berlin Database of Emotional Speech. Available from: http://www.elra.info/en/catalogues/catalogue-language-resources/