Information Sciences Letters
An International Journal

http://dx.doi.org/10.18576/isl/110401

An Improved Speech Emotion Classification Approach Based on Optimal Voiced Unit

Reda Elbarougy[1], *Noha M El-Badry*[2,*] *and Mona Nagy ElBedwehy*[3]

[1]Department of Information technology, Faculty of Computer and Information Sciences, Damietta University, New Damietta, Egypt
[2]Department of Mathematics, Faculty of Science, Damietta University, New Damietta, Egypt
[3]Department of Computer Science, Faculty of Computer and Information Sciences, Damietta University, New Damietta, Egypt

Received: 23 Dec. 2021, Revised: 12 Mar. 2022, Accepted: 29 Mar. 2022
Published online: 1 Jul. 2022

Abstract: Emotional speech recognition (ESR) has significant role in human-computer interaction. ESR methodology involves audio segmentation for selecting units to analyze, extract features relevant to emotion, and finally perform a classification process. Previous research assumed that a single utterance was the unit of analysis. They believed that the emotional state remained constant during the utterance, even though the emotional state could change over time, even within a single utterance. As a result, using an utterance as a single unit is ineffective for this purpose. The study's goal is to discover a new voiced unit that can be utilized to improve ESR accuracy. Several voiced units based on voiced segments were investigated. To determine the best-voiced unit, each unit is evaluated using an ESR based on a support vector machine classifier. The proposed method was validated using three datasets: EMO-DB, EMOVO, and SAVEE. Experimental results revealed that a voiced unit with five-voiced segments has the highest recognition rate. The emotional state of the overall utterance is decided by a majority vote of its parts' emotional states. The proposed method outperforms the traditional method in terms of classification outcomes. EMO-DB, EMOVO, and SAVEE improve their recognition rates by 12%, 27%, and 23%, respectively.

Keywords: Acoustic Features Extraction, Discriminative Features, Speech Emotion Recognition, Voiced Segments, Unvoiced Segments

1 Introduction

Sentiment analysis, considered one among the foremost important methods for analyzing communication in the real-world, is a kind of classification task to extract emotion from language. Many efforts have been made to enable machines to know the human emotions. Smart mobile devices, which incorporate speech recognition, can receive, and respond to voice commands through synthesized speech. Emotional Speech adopted for over 20 years [1,2,3,4,5] within the fields of human-machine interactions [6], the emotion integration in robotics [7], computer games [8], and the psychological assessment [9]. Although emotion recognition has been adopted throughout a broad way of everyday life, emotions involve subjective perceptions that make identifying human emotions a challenge. Although there are noteworthy progresses in recognizing emotions during a speech signal, in the field of the utterance-level emotional speech recognition (ULESR), there are still many challenges shall be solved. One of the challenges in ULESR found in dialogue system is that the same utterance can produce different emotions when it is in different contexts. [10] Because long utterances might demonstrate a variety of emotional states, the extracted low-level descriptive (LLD) acoustic features from such utterances are inconstant. As a result, using functional statistics like (mean, standard deviation) to get global statistics from the LLD for long utterances is unreliable. To emphasize the discriminative qualities of acoustic features over one unit, a standard emotion unit must be identified to produce a trustworthy statistic. As a result, sub-timing levels appear to be critical for enhancing ESR accuracy. A speech signal contains the unvoiced and voiced parts. When the vocal cords vibrate to pronounce vowels, the voiced segments are generated. Unlike the unvoiced segments, voiced segments manifest the periodic and prosodic signals. Unvoiced segments evince

* Corresponding author e-mail: noha_elbadry@du.edu.eg

the irregular signals have generated by the influence of the narrow vocal tract. Emotion is a crucial ingredient of the knowledge contained within the speech. Emotional information in a speech signal is represented in a variety of prosodic types and is especially contained within the voiced parts [11]. That is why we focused on the voiced parts of a speech in emotion recognition. The voiced parts of an utterance include the vowels that are very essential for ESR, because of the vowels are the richest parts with the emotional information [12]. Segmentation of voiced parts into its vowels is extremely challenging task and need either prior knowledge like the phoneme boundaries or using an ESR system to find these boundaries. On the other hand, segmenting into voiced segments is often easily done using voice activity detection with a really high performance. Accordingly, to extract the best and more related emotional information included within the vowel parts, and avoid the limitation of vowel segmentation, voiced segments are the best choice for voiced unit investigation. Obviously, the voiced segments are dynamic in terms of duration length. These dynamic properties of this unit are vital for capturing all changes within the emotional state during the utterance. During this study, it is assumed that one voiced segment cannot include more than one emotion, that is, during one voiced segment emotional state is fixed. It is hard to start out and end one emotional state in one voiced segment. However, the emotional state may persist/ continue for several consequences for voiced segments. It is not known how many voiced segments should be used to represent the optimal unit. To seek out the optimal unit, it is necessary to find unit with the fewest number of voiced segments that gives the best accuracy in emotion recognition. Therefore, the effect of including a different number of voiced segments in the proposed unit on ESR is investigated. The goal of this study is to find an acceptable segmentation method for dividing a speech signal into voiced units that represent emotions. Furthermore, to see if this unit can be used to improve ESR accuracy. The derived LLDs features from this unit are more consistent. As a result, extracting the global feature using some functional leads in more expressive features. Consequently, the overall emotional state of any utterance based on the emotional states of its constituents will be improved.

2 Speech Material

To improve the generalization capacities of the study's results, three different data sets (EMO-DB, SAVEE, and EMOVO) were used. One of the foremost widely used datasets for ESR is the Berlin Emotional Speech Database (EMO-DB). It is a free emotional dataset from Germany. The Institute for Communication Sciences, Technical University of Berlin, Germany developed EMO-DB. EMO-DB is a simulated dataset consisting of ten German sentences, five short sentences, and five long sentences. Ten professional speakers (Half of them are male and the other half are female) participated in data recording to create the dataset. Each one of the speakers had ten sentences, five long and five shorts, with different emotions. The EMO-DB dataset consists of seven emotions: *anger, sadness, boredom, happiness, anxiety, disgust*, and *neutral*. The data was recorded at a 48-kHz rate then down-sampled to 16-kHz [13]. The distribution of EMO-DB emotional states is shown in Fig 1.

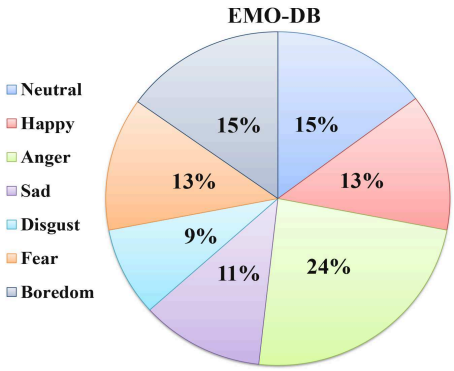

Fig. 1: The distribution of emotional states in the EMO-DB dataset

EMOVO is the first Italian language database for emotional speech. Six actors, three males and three females experienced, were summoned and asked to recite fourteen sentences representing six basic emotional states: disgust, sadness, fear, anger, joy, surprise, and neutral state. The recordings were made using appropriate professional instruments in the laboratories of the Fondazione Ugo Bordoni. The recordings were performed at a sampling frequency of 48 kHz, 16-bit stereo, and wav format [14, 15]. The distribution of EMOVO emotional states is shown in Fig 2.

The Surrey Audio-Visual Expressed Emotion (SAVEE) database is one of the foremost famous emotion datasets. SAVEE is widely used for developing an automatic emotion recognition system that was recorded as a prerequisite for this purpose. The SAVEE dataset consists of recordings for four native English male speakers' postgraduate students and researchers at the University of Surrey aged between 27 and 31 years. The total size of the SAVEE dataset is 480 utterances. Emotion has been described psychologically in seven discrete categories: anger, disgust, fear, happiness, sadness, and surprise and neutral [16]. The distribution of SAVEE emotional states is shown in Fig 3. The distribution of the datasets utilized within the study is given in Table 1.

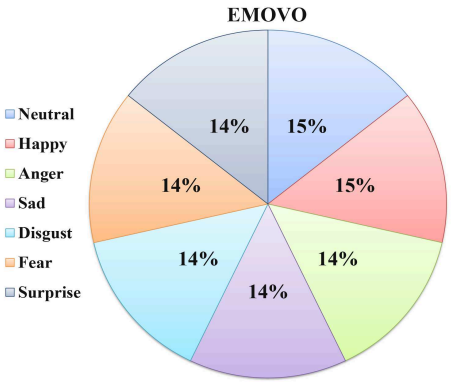

Fig. 2: The distribution of emotional states in the EMOVO dataset

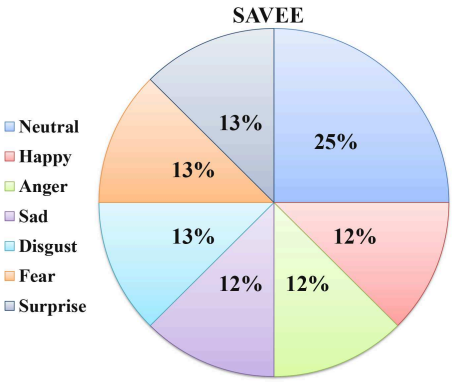

Fig. 3: The distribution of emotional states in the SAVEE dataset

Table 1. Distribution of speech recordings recordings utilized in the study.

Language	**EMO-DB (German)**	**EMOVO (Italian)**	**SAVEE (English)**
Neutral	79	84	120
Happy	71	84	60
Anger	127	84	60
Sad	62	84	60
Disgust	46	84	60
Fear	69	84	60
Boredom	81	-	-
Surprise	-	84	60
Total	**535**	**588**	**480**

3 Proposed Voiced Unit

This study assumes that a voiced unit within the voiced segments should be investigated. Subsequently, the method for segmentation of speech utterance into its voiced units is proposed in this section. These segments comprise F_0 information that is generally used to represent a speaker's emotional state. The most noteworthy segments of the emotional utterance are the voiced segments that include vowels which are very important for ESR, since vowels are the richest part with emotional information [17,18]. Vowels Segmentation is a very challenging task and requires either prior knowledge such as the phoneme boundaries or the use of the ASR system to determine these boundaries. On the other hand, segmentation into voiced segments can be easily performed using Voice Activity Detection (VAD) with very high performance [19,20]. As a result, to preserve the rich emotional information embedded in the vowel parts, and to avoid the limitation of vowel segmentation, voiced segments are the best candidates for voiced unit investigation. The segmentation into voiced segments is performed using the STRAIGHT software [21]. The algorithm used for this segmentation is based solely on acoustic information which makes it easy to be re-implemented in real time. Suppose the utterance U_i is segmented into its voiced segments using the following algorithm, the output of the segmentation process is the waveforms of all voiced segments which can be written as:

$$V(U_i) = \{V_{ij}, j = 1 : M_i\} \quad (1)$$

where (i) is the utterance index, V represents the sequence of all voiced segments for utterance U_i , V_{ij} is the j^{th} voiced segment, and M_i is the number of voiced segments in this utterance. Algorithm 1 is used to segment an utterance U_i into its voiced segments.

Algorithm 1. (Segmentation of speech utterance U into its voiced segments)

Segmentation of speech utterance algorithm
Input: An emotional speech utterance (U_i)
Output: The voiced segments $V_1, V_2, \ldots, V_N$
(1)Calculate the fundamental frequency F_0 for the speech utterance using STRAIGHT.
(2)For each frame in the utterance
If the value of F_0 is greater than zero, then
Label (frame) = voiced
Else
Label (frame) = Unvoiced F_0=Nan%
End if
(3) Determine the end points of the voiced segments (4) Use the end points time to obtain signal for each voiced segment: $V_{i1}, V_{i2}, \ldots, V_{iM_i}$

Voiced segments are dynamic in terms of duration length. These dynamic properties of this unit are very significant for capturing all changes in the emotional state during the utterance. Moreover, it is very rare for one voiced segment to include more than one emotion, which means that during one segment the emotional state is fixed. It is hard to start and end one emotional state in one voiced segment. However, the emotional state may persist for several consequences voiced segments.

We are not aware of how many voiced segments should be used to represent the optimal unit. To find the

optimal unit, it is necessary to find a unit with the fewest number of voiced segments that provides the best emotion recognition accuracy. Therefore, the effect of including a different number of voiced segments in the proposed unit on ESR is investigated. Figure 4 shows the process of segmentation into voiced units. First, the utterance is segmented into its voiced segments using the F_0 information extracted by the STRAIGHT software. Then, by applying the segmentation into units, this process combines several voiced segments to coordinate one voiced unit as explained in the rest of this section.

Thus, we define the voiced unit in terms of number of voiced segments. For example, voiced unit 1($VU^{(1)}$) is the method that segments an utterance into units/segments comprising one voiced segment as given by:

$$VU^{(1)}(U_i) = \{S_{ij} = V_{ij}, j = 1 : M_i\} \quad (2)$$

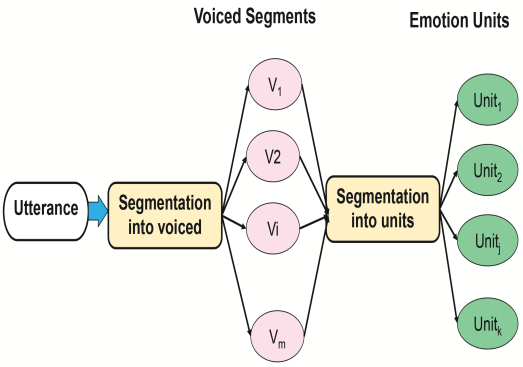

Fig. 4: Steps of segmentation into voiced units

where V_{ij}, M_i are as explained in Eq. (1) and S_{ij} is the j^{th} unit of utterance U_i. These units are simply the original voiced segments and there is no overlap between these units.

The second type is voiced unit 2 $\left(VU^{(2)}\right)$ that segments utterance into units containing two consequence voiced segments in each unit as given by:

$$VU^{(2)}(U_i) = \left\{ S_{ij} = \bigcup_{l=j}^{l=j+1} V_{i1}, j = 1 : M_i - 1 \right\} \quad (3)$$

This method definition depends on the use of a new windowed of the speech, using fixed length windows for two successive voiced segments with one voiced segment overlap. In general, the definition of voiced unit k $\left((VU^{(k)}\right)$ is

$$VU^{(k)}(U_i) = \left\{ S_{ij} = \bigcup_{l=j}^{l=j+k-1} V_{i1}, j = 1 : M_i - k + 1 \right\} \quad (4)$$

$VU^{(k)}$ segments the utterance U_i into units consisting of k voiced segments. From the above definition, it is clear that number of voiced units in one utterance depends on both the number of voiced segments and the type of unit representation. Figure 5 shows an example of one utterance segmentation with 6 voiced segments using segmentation method $VU^{(2)}$.

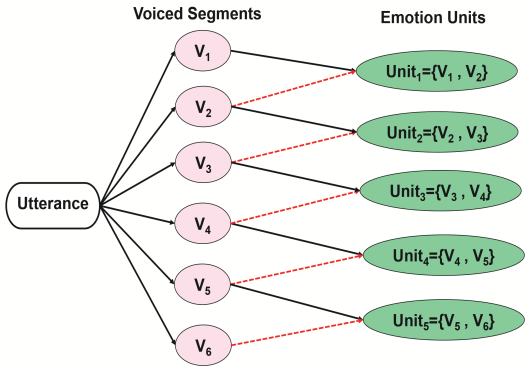

Fig. 5: Example of segmentation using $VU^{(2)}$ method for one utterance that have 6 voiced segments

Looking at the emotional speech database $DB = \{U_i, i = 1 : N\}$ of N emotional utterances and using different voiced unit segmentations on DB, different datasets of voiced units are obtained. For example, when applying $VU^{(k)}$ which includes k voiced segments, we obtain the following dataset:

$$VU^{(k)}(DB) = \left\{ S_{ij} = \bigcup_{l=j}^{l=j+k-1} V_{i1}, i = 1 : N, j = 1 : M_i - k + 1 \right\} \quad (5)$$

Units that have been obtained using this method retail have the same number of voiced segments. The number of units using this segmentation method is $\sum_{i=1}^{N} M_i$ where M_i is the number of voiced segments in utterance U_i.

To find the optimal voiced unit, the impact of including different number of voiced segments in the proposed unit on ESR is investigated. The unit that produces the highest recognition accuracy for the ESR system is the optimal unit.

In this study, the investigation for voiced unit is based on the categorical representation of emotion. Thus, traditional problem statement for emotion recognition based on this representation is reformulated according to the concept of voiced unit. Traditionally, ESR based on the categorical representation using utterance as a unit can be defined as follows: given a dataset of emotional speech utterances, each utterance is labeled with one class. The emotional categories of all utterances are given

by the following sequence:

$$C_{utterance} = \{C_i, i = 1 : N\} \quad (6)$$

where $C_{utterance}$ is the values of the emotional state for all utterance $\{U_i, i = 1 : N\}$. The conventional task of ESR is how to construct and train the ESR system to predict the emotional state for a new utterance.

The conventional approach can be reformulated using the voiced unit concept as given by Eq.(5). Since the label of each unit is not given in the original dataset, therefore, it is assumed that each voiced unit S_{ij} has the label of the utterance U_i which belong to. As a result, the emotional categories of all utterances are given by the following sequence:

$$C_{unit} = \{C(S_{ij}) = C_i, i = 1 : N, j = 1 : M_i - k + 1\} \quad (7)$$

Where C_{unit} is the values of the emotional state for all voiced segments in the unit S_{ij}, i, j and k are defined as in Eq.(5).

Subsequently, the new definition of the emotion recognition problem is as follow; given a dataset of voiced units as defined by Eq. (5) and the emotion classes for all units as given by Eq.(7), how to predict the emotion class for a new unit S_{ij} for which the system has not trained. Since one utterance contains a number of voiced units, therefore, predicting the emotional state of each unit is considered a continuous tracking of emotional states during one utterance.

Therefore, the proposed ESR system that was used to evaluate the impact of each voiced unit type is explained in detail as in Section 4.1. In addition, the traditional problem statement for classifying emotions is reformulated according to the concept of voiced unit as in Section 4.2.

4 Speech Emotion Recognition System

The proposed system for detecting emotional states consists of two stages, the training stage and therefore the testing stage. Within, utterance is segmented into its units during the training stage.

In the training stage, utterance is segmented into its units as described in Section 5. Then the acoustic features are extracted from each unit. The final step is to train the proposed classifier to know the relationship between the acoustic features extracted from the units and the emotional state of these units.

Furthermore, in the testing stage, the trained system is used to predict the emotional state of the new utterance. This stage includes 5 steps: the first step is used to segment the input utterance into its voiced segments, and then the resulting voiced segments are combined using Eq. (4) to constitute the voiced units as described in the previous section.

After that, the acoustic features have extracted from each voiced unit. In addition, these features are used as inputs to the trained SVM classifier to predict the emotional state of each voiced unit. Figure 6 shows the proposed method for predicting the emotional state for each unit. Moreover, the predicted labels for all units are used to determine the general emotional state of the entire utterance using the majority vote.

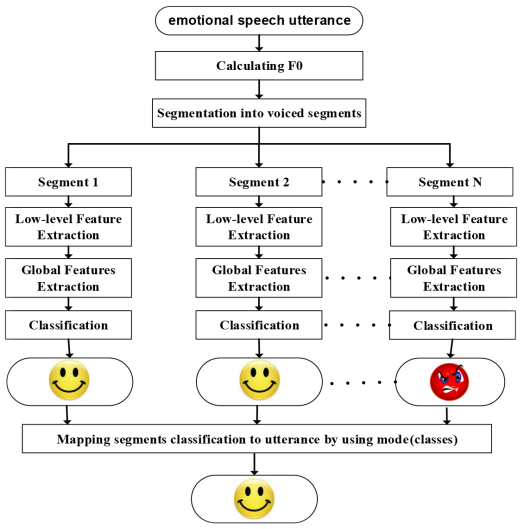

Fig. 6: Proposed ESR system using voiced unit concept

4.1 Datasets Used for The Experiment

To evaluate the proposed system, we use the utterance with at least 5 voiced segments. Table 2 shows the number of utterances that include at least 5 voiced segments in the three databases presented in Section 2. The distribution of speech recordings utilized in the study is given in Table 3.

Table 2. The number of utterances in the used datasets

	EMO	EMOVO	SAVE
Utterance (#)	512	522	429

Table 3. Distribution of speech recordings utilized in the study

Language	EMO-DB (German)	EMOVO (Italian)	SAVEE (English)
Neutral	74	73	108
Happy	69	78	54
Anger	125	81	53
Sad	62	66	54
Disgust	45	69	55
Fear	63	78	49
Boredom	74	0	0
Surprise	0	77	56
Total	512	522	429

Table 4 shows the number of units after applying 5 segmentation methods; $VU^{(k)} : k = 1 \ldots 5$ for the

utterances in Table 2 for the three databases. For example, the first row in Table 4 includes the number of segmented units using the segmentation method $VU^{(1)}$, the 512 utterances of EMO-DB are segmented into 5,541 units consisting of one voiced segment.

Table 4. Utterance distribution for the selected database for different segmentation methods

Segmentation methods		$VU^{(1)}$	$VU^{(2)}$	$VU^{(3)}$	$VU^{(4)}$	$VU^{(5)}$
Data	EMO	5.541	2.647	1.185	906	733
	EMOVO	6.497	3.125	1.425	1,085	883
	SAVEE	6.674	3.229	1.512	1,163	937

In order to compare the proposed segmentation method into voiced unit based on voiced segments, we segment the three databases using fixed duration units.

As the first variant apart from the utterance-level features, we consider fixed time units. The database is split into fixed units first, and then classified. Table 5 shows the obtained results and the number of units for the five diverse splitting unit lengths chosen.

Table 5. Utterance distribution for the selected databases for different fixed segmentation methods

		$FU^{(1)}$	$FU^{(2)}$	$FU^{(3)}$	$FU^{(4)}$	$FU^{(5)}$
Duration $\geq$		0.25 sec	0.5 sec	1 sec	1.25 sec	1.5 sec
Data	EMO	3.972	4.294	3.869	3.363	2.852
	EMOVO	3.883	4.519	4.061	3.545	3.023
	SAVEE	3.761	4.140	3.830	3.428	3.010

4.2 Features Extraction

Feature extraction for ESR generally involves four stages as illustrated in Fig. 7. There are stages for framing, windowing, low-level feature extraction, and global feature calculation. To minimize effort or improve classification accuracy, signal processing methods are utilized for pre- and post-processing.

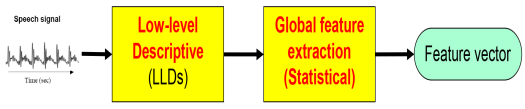

Fig. 7: The Block Diagram for feature extracting

The emobase2010 configuration is profusely used in ESR. OpenSMILE v2.1.0 software [22] During the feature extraction stage, the emobase2010 configuration file was used and 1582 features were obtained from each speech record. Table 6 lists the specific features.

The LLDs audio features in terms of speech parameters presented in the first column of Table 6 are extracted from each frame in each unit. Moreover, the delta for each LLDs is also extracted.

Table 6. The used acoustic feature (Statistical Functional Groups (A, B, C) are specified in Table 7)

(LLDs)	#	Δ	Global-Features	#
PCM Loudness	1	✓	A, B	42
MFCC [0 -14]	15	✓	A, B	630
log MFB [0-7]	8	✓	A, B	336
LSP Freq. [0-7]	8	✓	A, B	336
F_0 envelope	1	✓	A, B	42
F_0 by SHS	1	✓	A, B	42
Voicing probability	1	✓	A	38
Jitter Local	1	✓	A	38
Jitter DDP	1	✓	A	38
Shimmer Local	1	✓	A	38
F_0 by SHS	-	×	C	2
	38			1582

Note that the abbreviation DDP, LSP, MFB, and SHS is defined as follows:

–(DDP) stands for Difference of Difference of periods,
–(LSP) is the abbreviation of Line Spectral Pairs, (MFB) is Mel Frequency Band
–(SHS) is sub-Harmonic sum

Then, apply the statistical functional sets (A, B, C) shown in the 4^{th} column that are defined in Table 7 to extract the global features form each unit.

Table 7. The used Statistical functions to calculate the global features

Set	#	Sets of Statistical functions
A	19	Position max/min Arithmetic average, standard deviation, skewness, kurtosis Linear regression coefficient 1 / 2 Quadratic & absolute linear regression error Quartile 1 / 2 / 3 Quartile range 1-2 / 2-3 / 1-3 Percentile 99 Up-level time 75/90
B	2	Percentile 1, percentile range 0-1
C	2	Onsets number, Duration

4.3 The Used Classifier

To determine the optimal voiced unit, the impact of each voiced unit on ESR is investigated. Thus, SVM classifier

was used for evaluation that is based on statistical learning theory. Given the training data: $D = \{\{x_i, y_i\}, i = 1 : N\}$, where $x_i \in R^d$ represents the sample features in the d-dimensional space, $y_i \in \{-1, +1\}$ represents the class labels, and N is the number of samples. For any sample x_i, in the training set, let a linear classifier characterized by the set of pairings (w, b) satisfies the following optimum hyperplane inequalities:

$$\begin{cases} w.x_i + b \geq 1 - \xi_i & \text{if } y_i = +1 \\ w.x_i + b \leq -1 + \xi_i & \text{if } y_i = -1 \end{cases} \quad (8)$$

Where w stands for the weight vector, the tendency value is b, and ξ_i denotes a positive artificial variable. The SVM classifier is based on the artificial variable ξ_i as follows: If $\xi_i = 0$, then the sample x_i is successfully classified. If ξ_i is in the range;$0 < \xi_i < 1$, then x_i is also properly classified, but its position is among the extreme planes. It is incorrect classified when $\xi_i > 1$.

The kernel function is used to classify the data in a higher dimension when the two-class problem cannot be linearly separated. The standard Kernel functions include the "*linear*," "*polynomial*," "*radial basis*," and "*sigmoid*" functions. SVM is generally used to solve two class problems. For multiple classification, One-Against-Rest, One-Against-One, and Multi-Class Ranking techniques can be used. In this study, the RBF kernel function was used in the SVM classifier. The acoustic features obtained in section 4.3 are reflected in sample x_i in this research. Without any feature selection or reduction, SVM is applied to the best of all 1582 acoustic properties. The suggested kernel function is the radial basis kernel function. To avoid measurement unit differences between the features sets obtained, all values are normalized between 0 and 1.

4.4 Emotion Classification Results

To identify the emotional state of the whole utterance, we first utilize the SVM classifier to predict the emotional state of the utterance's units, which are used to determine the general emotional state of the entire utterance using the majority vote. As a result, we provide details of these two stages in the subsections below.

4.4.1 Voiced Unit Classification Results

This section illustrates the proposed method's first stage accuracy. Five SVM classifiers were trained for each database using the five databases of voiced units presented in Table 5 based on the proposed voiced unit segmentation. We extract 1582 features for each unit using the OpenSMILE software. The extracted acoustic features were used to train and test the SVM classifier. The block diagram for unit classification is shown in Fig. 8.

The classification rate for the units of the utterance "11b03Fc" from EMO-DB, for example, is shown in Figs. 9-11 based on segmentation using $VU^{(1)}, VU^{(2)}$ and $VU^{(5)}$ respectively. The emotional state of used utterance is happy; therefore, each original unit is labeled by happy which corresponds to class 2. The predicted classes for all units were misclassified.

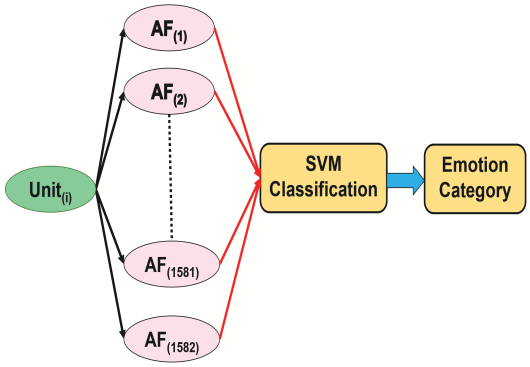

Fig. 8: The Block Diagram for unit classification

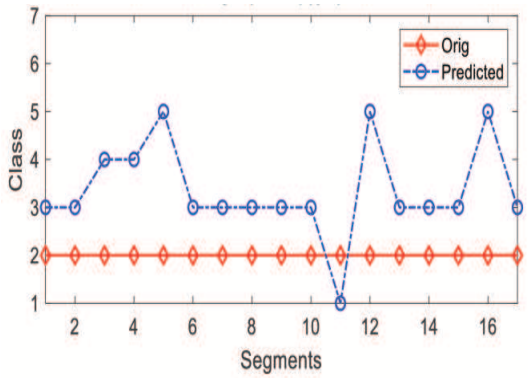

Fig. 9: Unit classification using segmentation $VU^{(1)}$

However, applying the segmentation technique $VU^{(2)}$ as indicated in Fig. 10, the projected classes' accuracy is little improved, as shown in Fig. 9. When the segmentation technique $VU^{(5)}$ was used, the classification accuracy increased to about 78 %, as shown in Fig. 11.

The proposed approach for ESR is used to determine which voiced unit from the five types of units $VU^{(1)}, VU^{(2)}, VU^{(3)}, VU^{(4)}$ and $VU^{(5)}$ is the optimum. These types are used to segment each database separately. For each database, five emotional unit datasets are

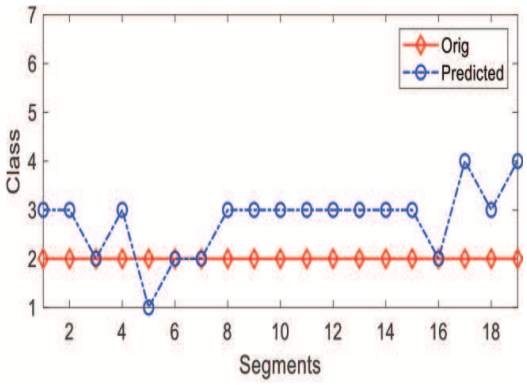

Fig. 10: Unit classification using segmentation $VU^{(2)}$

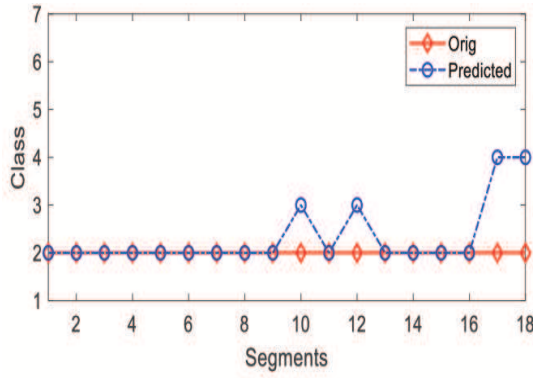

Fig. 11: Unit classification using segmentation $VU^{(5)}$

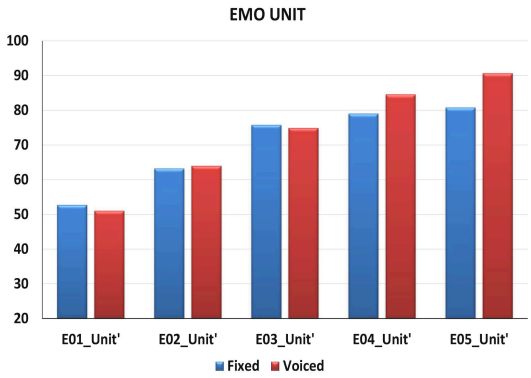

Fig. 12: Unit recognition rate for five types of segmentation methods for EMO-DB

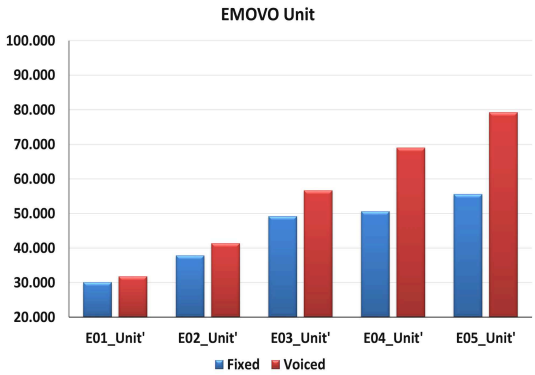

Fig. 13: Unit recognition rate for five types of segmentation methods for EMOVO

obtained. . The five data sets were used separately to train and test the proposed SER system using 10-fold cross-validation to measure the effect of each unit on the recognition rate of emotional state. Acoustic features for each dataset are the system's inputs, and emotional class is the system's output. The suggested voiced segmentation methods were compared against the five types of fixed time segmentation units $FU^{(1)}, FU^{(2)}, FU^{(3)}, FU^{(4)}$ and $FU^{(5)}$. Figures 12-14 show the classification rate for all datasets for the databases EMO-DB, EMOVO, and SAVEE, respectively.

These figures show that the voiced unit $VU^{(5)}$ technique attained the highest recognition rate. As a result, this segmentation approach is considered as the optimal unit. To analyze speech in continuous tracing for identifying emotional changes, it is necessary to segment the utterance into units with a window of five consequence voiced segments with overlap of four voiced segments, according to the findings of this research. The classification accuracy for unit recognition accuracy for the three databases for the two methods of segmentation is summarized in Fig. 15.

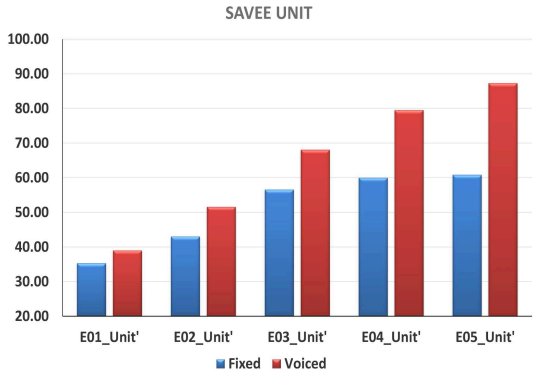

Fig. 14: Unit recognition rate for five types of segmentation methods for SAVEE

4.4.2 Results of Utterance Classification

Furthermore, compared to the previous study, the label of the whole utterance is predicted using a majority vote of the predicted labels of its units. As a result, the proposed

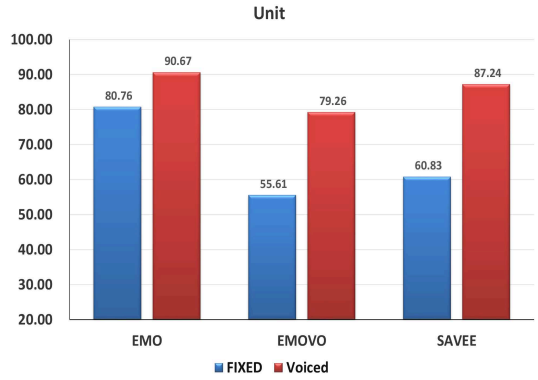

Fig. 15: Comparing the segmentation methods voiced and fixed units for the three databases

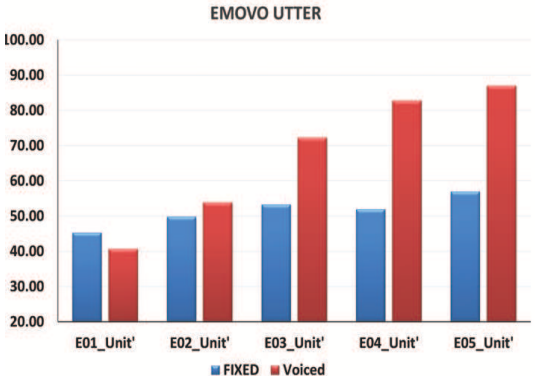

Fig. 17: Utterance-level recognition rate for five types of segmentation methods for EMOVO

system is able not only to detect changes in the emotional state during utterance but also to predict the overall emotional state in the whole utterance. Figures 16-18 show the results of emotion classification for the whole utterance using the proposed technique and the conventional method for the three datasets.

The results reveal that increasing the segment duration improves the results for both fixed and voiced units. However, the voiced unit is superior to the fixed time unit segmentation method.

To clarify the effectiveness of the proposed technique, the obtained results were compared with previous studies that used the same corpus. The traditional approaches that use the whole utterance as one unit are compared to the proposed method based on the voiced segments.

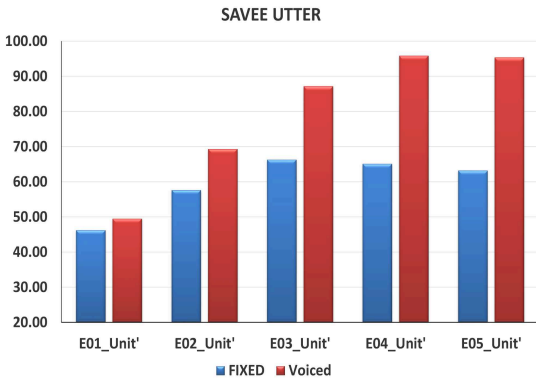

Fig. 18: Utterance-level recognition rate for five types of segmentation methods for SAVEE

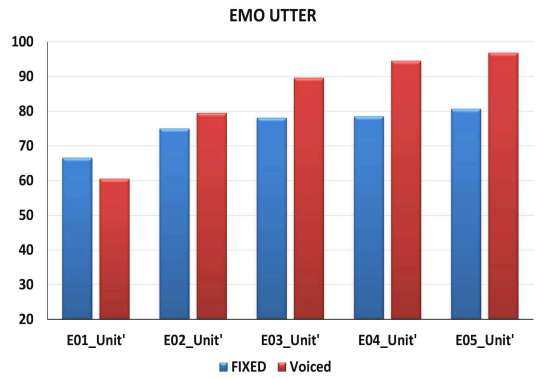

Fig. 16: Utterance-level recognition rate for five types of segmentation methods for EMO-DB

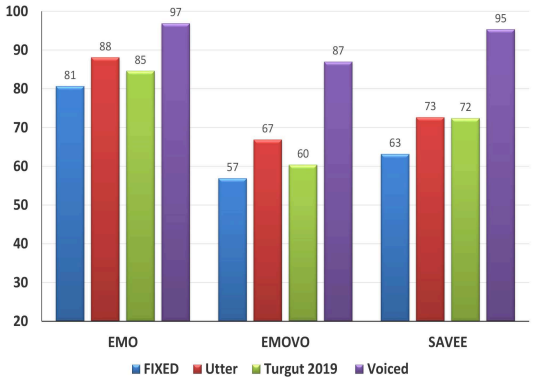

Fig. 19: Comparing the proposed method with the traditional methods

It is found that the proposed method is superior the traditional method for emotion classification described by Turgut Özseven in (2019) [23]. The improvement of the proposed method is 12% from 85% to 97% for EMO database. Moreover, great improvement for the EMOVO and SAVEE databases, with improvements of 27% and 23%, respectively. Furthermore, the proposed method outperforms the traditional method even with applying feature selection is used with an improvement of 2.3%

from 82% to 84.3%. The proposed method for emotion classification is characterized by the ability to accurately predict the emotion category using the voiced units, with a very high emotion classification rate.

5 Conclusions

A segmentation method for dividing a speech utterance into voiced units has been proposed to improve the classification accuracy of ESR. The following candidates for the voiced unit have been proposed: $VU^{(1)}$,$VU^{(2)}$,$VU^{(3)}$,$VU^{(4)}$ and $VU^{(5)}$. Each voiced unit is determined by the number of voiced segments, the voiced unit of type $VU^{(i)}$ has i voiced segments. The SVM classifier is used to evaluate the impact of each voiced unit on the ESR.

To train the classifier, 1582 features were extracted for each unit using OpenSMILE software, and the experimental results reveal that the $EU^{(5)}$achieved the highest recognition rate. Thus, the predicted classes of units were used to predict the emotional state of the whole utterance using majority voting. Three datasets of emotional speech were used to validate the proposed method: EMO-DB, EMOVO, and SAVEE. The classification results using the proposed method outperforms the conventional method, and the improvements in recognition accuracy were 12%, 27% and 23% EMO-DB, EMOVO, and SAVEE, respectively.

Acknowledgment

This paper was produced with the financial support of the Academy of Scientific Research and Technology of Egypt; ScienceUP/GradeUp initiative: Grant Agreement No (6661). Its contents are the sole responsibility of the authors and do not necessarily reflect the views of the Academy of Scientific Research and Technology.

Competing interests: The authors declare that they have no competing interests.

References

[1] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, M. Mahjoub and C. Cleder, *Automatic Speech Emotion Recognition Using Machine Learning*, In Social Media and Machine Learning; IntechOpen, 2019.

[2] B. Schuller, Speech emotion recognition: Two decades in a nutshell, benchmarks, and ongoing trends, *Communications of the ACM*, **61(5)**,90-99, 2018.

[3] T. Jayasankar, K. Vinothkumar and A. Vijayaselvi, Automatic Gender Identification in Speech Recognition by Genetic Algorithm, *Applied Mathematics & Information Sciences*, **11(3)**, 907-913, 2017.

[4] A. Shoiynbek, K. Kozhakhmet, N. Sultanova and R. Zhumaliyeva, The Robust Spectral Audio Features for Speech Emotion Recognition, *Applied Mathematics and Information Sciences*, **13(5)**, 867-870, 2019.

[5] P. Sujatha and M. Radhakrishnan, Mouth Segmentation Using Coordinate-Based Method for the Improvement of Visual Speech Recognition, *Applied Mathematics and Information Sciences*, **12(4)**, 891-897, 2018.

[6] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. Taylor, Emotion recognition in human-computer interaction, *IEEE Signal Processing Magazine*, **18(1)**, 32-80, 2001.

[7] X. Huahu, G. Jue and Y. Jian, *Application of Speech Emotion Recognition in Intelligent Household Robot*, in In Proceedings of the 2010 International Conference on Artificial Intelligence and Computational Intelligence, Sanya, China, 2010.

[8] M. Szwoch and W. Szwoch, Emotion recognition for affect-aware video games, *In Image Processing & Communications Challenges*, **6**, 227-236, 2015.

[9] L. Low, N. Maddage, M. Lech, L. Sheeber and N. Allen, Detection of Clinical Depression in Adolescents' Speech During Family Interactions, *IEEE Transactions on Biomedical Engineering*, **58(3)**574-586, 2010.

[10] Q. Li, C. Wu, Z. Wang and K. Zheng, Hierarchical Transformer Network for Utterance-Level Emotion Recognition, *Applied Sciences*, **10(4447)**, 1-13, 2020.

[11] R. Elbarougy, Speech Emotion Recognition based on Voiced Emotion unit, *International Journal of Computer Applications*, **178(47)**, 22-28, 2019.

[12] F. Ringeval and M. Chetouani, *A vowel based approach for acted emotion recognition*, in Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, Brisbane, Australia, 2008.

[13] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, *A database of German emotional speech*, in Proceedings of the Ninth European Conference on Speech Communication and Technology, Lisboa, Portugal, 2005.

[14] I. Iadarola, *EMOVO, database di parlato emotivo per l'italiano*, in Atti del 4° Convegno Nazionale dell'Associazione Italiana di Scienze della Voce, Arcavacata di Rende (CS), 2007.

[15] G. Costantini, I. Iaderola, A. Paoloni and M. Todisco, *Emovo corpus: an italian emotional speech database*, in Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), Reykjavik, Iceland, 2014.

[16] P. Jackson and S. Haq, *Surrey Audio-Visual Expressed Emotion (SAVEE) Database*, 2014.

[17] B. Vlasenko, D. Philippou-Ḧubner, D. Prylipko, R. B̈ock, I. Siegert and A. Wendemuth, *Vowels formants analysis allows straightforward detection of high arousal emotions*, in IEEE International Conference on Multimedia and Expo (ICME), 2011.

[18] D. S. and S. Dandapat, Emotion Classification Using Segmentation of Vowel-Like and Non-Vowel-Like Regions, *IEEE Transactions on Affective Computing*, **10(3)**, 360-373, 2017.

[19] M. H. Moattar and M. M. Homayounpour, *A simple but efficient real-time Voice Activity Detection algorithm*, in 17th European Signal Processing Conference, Glasgow, UK, 2009.

[20] M. H. Moattar, M. M. Homayounpour and N. K. Kalantari, *A new approach for robust realtime Voice Activity Detection using spectral pattern*, in IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, 2010..

[21] H. Kawahara, I. Masuda-Katsuse and A. Cheveign, Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F_0 extraction: Possible role of a repetitive structure in sounds, *Speech Communication*, **27(3-4)**, 187-207, 1999.

[22] F. Eyben, F. Weninger, F. Gross and B. Schuller, *Recent developments in openSMILE, the munich open-source multimedia feature extractor*,in Proceedings of the 21st ACM international conference on Multimedia, 2013.

[23] T. Özseven, A novel feature selection method for speech emotion recognition, *Applied Acoustics*, **146**, 320-326, 2019.