Speech Emotion Recognition based on Voiced Emotion Unit

Reda Elbarougy
Department of Computer Science
Faculty of Computer and Information Sciences, Damietta University
New Damietta, Egypt

ABSTRACT
Speech emotion recognition (SER) system is becoming a very important tool for human-computer interaction. Previous studies in SER have been focused on utterance as one unit. They assumed that the emotional state is fixed during the utterance, although, the emotional state may change during the time even in one utterance. Therefore, using utterance as one unit is not suitable for this purpose especially for long utterances. The ultimate goal of this study is to find a novel emotion unit that can be used to improve SER accuracy. Therefore, different emotion units defined based on voiced segments are investigated. To find the optimal emotion unit, SER system based on support vector machine (SVM) classifier is used to evaluate each unit. The classification rate is used as a metric for the evaluation. To validate the proposed method, the Berlin database of emotional speech EMO-DB is used. The experimental results revealed that emotion unit that contains four voiced segments gives the highest recognition rate for SER. Moreover, the final emotional state of the whole utterance is determined by majority voting of emotional states of its units. It is found that the performance of the proposed method using voiced related emotion unit outperforms the conventional method using utterance as one unit.

General Terms
Pattern Recognition, Machine learning, speech processing.

Keywords
Acoustic features extraction, discriminative features, speech emotion recognition, voiced segments, unvoiced segments.

1. INTRODUCTION
Automatic speech emotion recognition (ASER) systems becoming a very important technology for human-computer interaction [1]. This technology can be embedded in many computer applications as well as in robots to make them sensitive to the user's emotional voice [2]. Methodology for ASER includes audio segmentation to find units for analysis, extraction of emotion-relevant features, and classification. The major challenges for real-time applications is audio segmentation into an appropriate unit for emotions [4, 5, 6]. Most approaches so far have dealt with utterances of acted emotions where the choice of unit is visibly just one utterance, a well-defined linguistic unit with no change of emotion within it. Form the literature, several indications originated from acted data recorded in the lab; for this data, beginning and end are given by using utterance as one unit [7, 8, 9]. However, in spontaneous speech this kind of observable unit does not exist. It is known that segmentation into utterances is not straight-forward moreover, the emotion is not constant over an utterance.

Linguistically motivated units such as utterance has many limitations for continuous emotion recognition for the following reasons: (1) it requires preprocessing by an automatic speech recognition (ASR) system to obtain unit boundaries, (2) time constrictions make it obligatory not to wait with ASR till the speaker has complete his/her whole turn, (3) emotion is dynamic and may changes during the utterance, especially in spontaneous speech, utterances' duration are much varied from very short to very long utterance. Therefore, long utterances may include different emotional states. Thus, the extracted low-level descriptive (LLD) acoustic features such as Mel-frequency cepstral coefficient (MFCC) from such utterances are inconstant because they demonstrating different emotional states. As a result, applying functional statistics such as (mean, standard deviation) to obtain global statistic form the LLD for long utterance are not reliable. Therefore, to emphasis the discriminative properties of acoustic features over one unit, it is needed to find a standard emotion unit to obtain a reliable statistic. Therefore, sub-timing levels seem to be very important for improving SER [4].

The goal of this study is to find an appropriate segmentation method that can be used for segmenting a speech signal into its emotion units which are representative for emotions. Moreover, to investigate the feasibility of this unit to improve SER accuracy. Finding an optimal emotion unit that include only one emotional state, make the extracted LLD feature from this unit more consistent. Thus, applying some functional to extract the global feature leads to more expressive features. Segmenting one utterance into its emotion units will help use to determine the emotional state of each unit. As a result, we can determine the going on emotional state in real-time in case of continuous speech. As well as to determine the overall emotional state of any utterance from the emotional states of its units.

2. SPEECH MATERIAL
The Berlin emotional speech database (EMO-DB) contains 535 utterances spoken by actors in seven emotional states: happy, anxious, bored, disgusted, angry, fearful, and neutral [10]. The utterances were recorded by ten actors, five of them were females and five were males with ten different sentences in German language. The 10 utterances can be divided into five long sentences (around 4s) and the other five are short sentences (nearly 1.5 s). These sentences were not equally distributed between the various emotional states, it is required to make balance for training purpose, therefore, an equal distribution of the four emotional states was used, 50 neutrals, 50 sad, 50 angry, and 50 happy. To sum up, 200 utterances were selected from the EMO database: 100 utterances were spoken by five females and the other 100 by five males divided equally between the four emotional states.

3. PROPOSED EMOTION UNIT

This section explains the proposed method for segmentation of speech utterance into its emotion units. This study assume that an emotion unit should be investigated within the voiced segments. These segments comprise F0 info that are generally used to represent emotional state of the speaker. The most significant segments of an emotional utterance are the voiced segments that include vowels which are very important for SER, due to vowels are the richest part with emotional information [11, 12, 13]. Vowels Segmentation is very challenging task and require either prior knowledge such as the phoneme boundaries or using an ASR system to determine these boundaries. On the other hand, segmentation into voiced segments can be easily done using voice activity detection (VAD) with a very high performance [14, 15]. As a result, to keep the rich emotional information included in the vowel parts, and avoid the limitation of vowel segmentation, voiced segments are the best candidates for emotion unit investigation.

The segmentation into voiced segments is done by using STRAIGHT software [16]. The algorithm used for this segmentation is based only on acoustic information that make it easy to be re-implemented in real-time. Suppose the utterance U_i is segmented into its voiced segments using this algorithm, the output of segmentation process are the waveforms of all voiced segments which can be written as follows:

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

where (i) is the utterance index, V represents the sequence of all voiced segments for utterance U_i, V_{ij} is the jth voiced segment, and M_i is the number of voiced segments in this utterance. For example, the result of segmentation of an utterance U_i into its voiced segments is shown in Figure 1. In this figure the used utterance consists of 6 voiced segments i.e. $M_i = 6$. Therefore, the segmentation yields 6 units as given by:

$$\mathrm{V}(U_i) = \{V_{i1}, V_{i2}, V_{i3}, V_{i4}, V_{i5}, V_{i6}\} \quad (2)$$

From this figure, it is clear that voiced segments are dynamic in terms of duration length. These dynamic properties of this unit is very important to capture all changes on 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 emotional state is fixed. It is difficult to start one emotional state and end it in one voiced segment. However, the emotional state may continue for several consequences voiced segments.

To answer the research question what is the optimal emotion unit? It is not known how many voiced segments should be used to represent the optimal unit. To find the optimal unit, it is necessary to find unit with minimum number of voiced segments that gives the best emotion recognition accuracy. Therefore, impact of including different number of voiced segments in the proposed unit on SER is investigated.

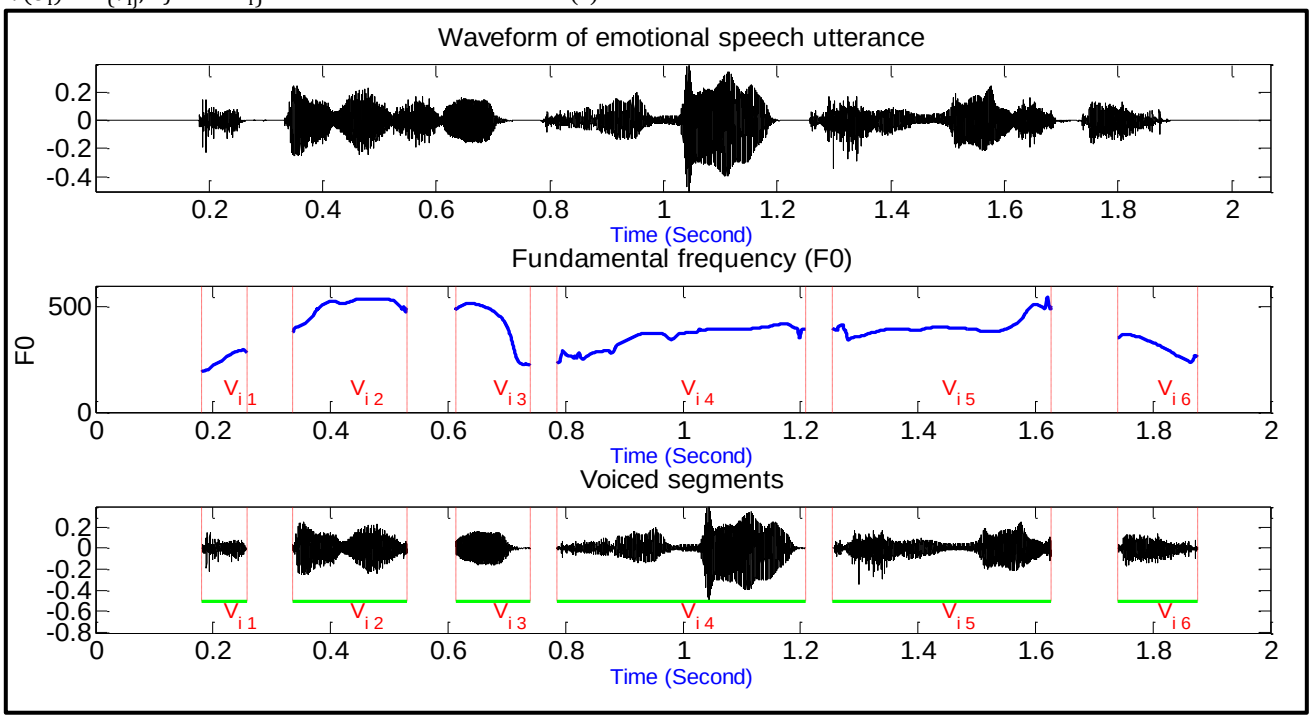

Figure 1: Segmentation of speech utterance U_i into its voiced segments V based on F0 information extracted using STRAIGHT software, i is an utterance index.

Thus, for example, what we call emotion unit 1 ($EU^{(1)}$) is the segmentation method that segments utterance into units/segments that include one voiced segment in each, and is defined by:

$$EU^{(1)}(U_i) = \{S_{ij} = V_{ij},\ j = 1{:}M_i\} \quad (3)$$

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

The second type of segmentation method is emotion unit 2 ($EU^{(2)}$) that segments utterance into units that contains two consequence voiced segments in each unit is given by

$$EU^{(2)}(U_i) = \left\{S_{ij} = \cup_{l=j}^{l=j+1} V_{il},\ j = 1{:}M_i - 1\right\} \quad (4)$$

The definition of this method is based on the use of a new windowed of the speech, using a windows of fixed length of two consequence voiced segments with overlap of one voiced segment. For example, the first unit using this representation is a sequence of 1st and 2nd voiced segments $S_{i1} = \{V_{i1}, V_{i2}\}$ and and $S_{i2} = \{V_{i2}, V_{i3}\}$ and so on.

Similarly, emotion unit 3 $(EU^{(3)})$ is a sequence of three voiced segments. This type has a windows of fixed length of three consequence voiced segments with overlap of two voiced segment as given by

$$EU^{(3)}(U_i) = \left\{S_{ij} = \cup_{l=j}^{l=j+2} V_{il},\ j = 1{:}M_i - 2\right\} \quad (5)$$

In a similar way, emotion unit 4 $(EU^{(4)})$ that consists of four voiced segments is given by this equation

$$EU^{(4)}(U_i) = \left\{S_{ij} = \cup_{l=j}^{l=j+3} V_{il},\ j = 1{:}M_i - 3\right\} \quad (6)$$

In general we can define emotion unit k $(EU^{(k)})$ by

$$EU^{(k)}(U_i) = \left\{S_{ij} = \cup_{l=j}^{l=j+k-1} V_{il},\ j = 1{:}M_i - k + 1\right\} \quad (7)$$

This representation segments the utterance U_i into units/segments which consists of k voiced segments.

From the above definition, it is clear that, the number of emotion units for one utterance depend on both the number of voiced segments and the type of unit representation.

4. EVALUATION OF THE PROPOSED EMOTION UNIT

This section introduces the proposed method for finding the optimal segmentation method. However, it is possible to extend our investigation for any segmentation methods defined by equation 7, i.e. for any unit that contain k voiced segments. However, most of utterances in the used database includes at least four voiced segments. Therefore, we limit our investigation for the optimal unit among the first four proposed segmentation methods $(EU^{(1)}, EU^{(2)}, EU^{(3)}, EU^{(4)})$ defined in the previous section by eq. 3-6. Appling these segmentation methods on EMO-DB introduced in section 4.2, four datasets of emotion units are obtained, namely, EU1-DB, EU2-DB, EU3-DB, and EU4-DB. To accomplish this task, the impact of the four methods on recognition rate is investigated. The unit that yields the highest recognition accuracy for SER system is selected as the optimal one. In this study, the investigation for emotion unit is based on the categorical representation of emotion. Therefore, the proposed SER system that used for evaluating the impact of each emotion unit type is explained in details as in section 4.1. In addition, the traditional problem statement for emotion classification is reformulated according the concept of emotion unit as in section 4.2.

4.1 Speech emotion recognition system

The proposed system for detecting the emotional state is shown in Fig 2. This system is composed of two stages: the training phase and the testing phase. In the first stage utterance is segmented into its emotion units as explained in section 3. Then acoustic features extracted from each emotion unit as introduced in section 4.3. The third step is implementing a feature selection method to use only the most discriminative acoustic features. It is assumed that each emotion unit has the same emotion class as the utterance they belong to. The final step is to train the proposed classifier to learn the relationship between acoustic features extracted from emotion unit and the emotional state of this unit.

The second stage is the testing phase; this stage is used to predict the emotional state of a new utterance using the trained SER system.

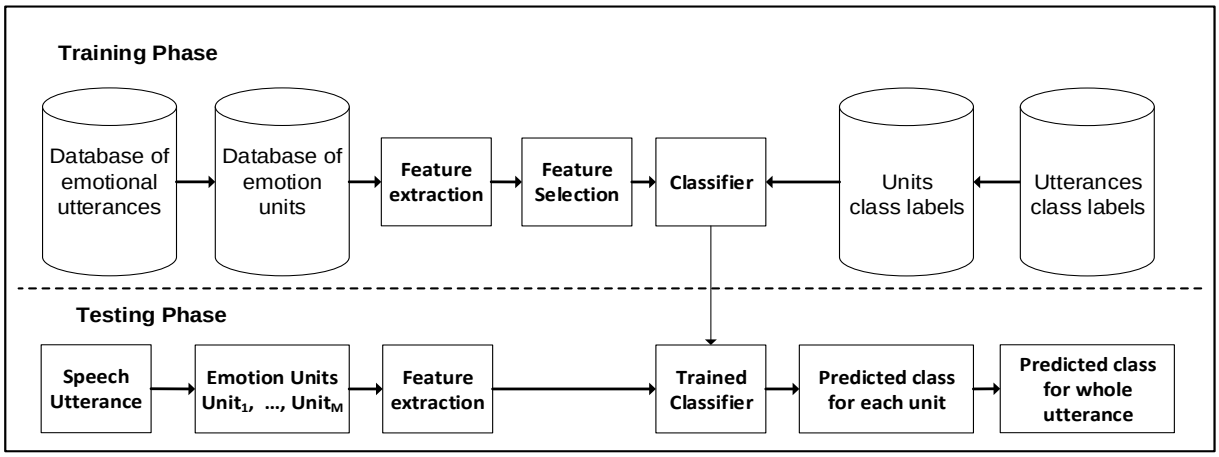

Figure 2: Proposed SER system using voiced emotion unit concept

This stage includes 5 steps: the first step is used to segment the input utterance into its emotion units. Then extract the selected acoustic features from each emotion unit. In addition, these features are used as inputs to the trained classifier to predict the emotional state of each emotion unit. Moreover, the predicted labels for all units are used to determine the overall emotional state of the whole utterance by using the majority voting.

4.2 Problem statement using emotion unit

Traditionally, SER based on categorical approach can be defined as follows: given a dataset of N emotional speech utterances: $U = \{U_i,\ i = 1{:}N\}$. Each utterance is labeled using the categorical approach. The emotional categories of all utterances are given by the following sequence $C = \{C_i,\ i = 1{:}N\}$, where C_i is the emotional state of the utterance U_i. The conventional task of SER is how to construct and train SER system that has the ability to classify each utterance into emotional state.

Using emotion unit concept, the conventional approach could be reformulated as follows, suppose for example, $EU^{(1)}$ is used for segmentation, the obtained dataset is EU1-DB, this dataset includes all units composed of one voiced segment.

$$\text{EU1} - \text{DB} = \{S_{\text{ij}} = V_{\text{ij}},\ i = 1 \rightarrow N,\ j = 1{:}\,M_{\text{i}}\} \quad (8)$$

Since the label of each unit is not given in the database, therefore, it is assumed that each emotion unit S_{ij} has the same label of the utterance U_{i} which belong to. As a result, the labels of all units in the used database is given by

$$C^{(1)} = \{C_{\text{ij}} = C_{\text{i}},\ i = 1 \rightarrow N,\ j = 1 \rightarrow M_{\text{i}}\} \quad (9)$$

where C_{i} is the emotion category of utterance U_{i} and C_{ij} is the category of unit number j from utterance U_{i}. Therefore, the new definition for emotion classification problem is as follows: given a dataset of emotion units EU1-DB and the emotion categories labels of all units $C^{(1)}$, how to predicted the emotion category label for a new unit S_{ij}.

The rest of this subsection explains the details of applying the proposed emotion unit segmentation methods introduced in the previous subsection. Tables 1-4 shows the detailed information for both utterance datasets which were selected from the original database and units' datasets that obtained by using segmentation method. Table 1, gives distribution of emotion classes for the original utterance that can be used in each segmentation method. For example, the first row indicate that all the 200 utterances can be segmented using the segmentation method $EU^{(1)}$, therefore the used dataset is the original database EU1-DB is equal to EMO-DB that contain the 200 utterances.

It is clear that all utterances include at least 3 voiced segments as shown in the first 3 rows: EU1-DB, EU2-DB, and EU3-DB. Row number four shows the utterances that can be segmented using $EU^{(4)}$, from this information, two utterances does not include four voiced segments and the other 198 utterances contains at least four voiced segments.

Table 1. Utterance distribution of emotion classes for the selected database for different segmentation methods

Segmentation method	Dataset	Utterance (#)	Emotion Classes			
			Neutral	Happiness	Anger	Sadness
$EU^{(1)}$	U1-DB	200	50	50	50	50
$EU^{(2)}$	U2-DB	200	50	50	50	50
$EU^{(3)}$	U3-DB	200	50	50	50	50
$EU^{(3)}$	U4-DB	198	49	50	49	50

Table 2, shows the duration statistics for the utterances presented in the above table. The minimum, maximum, mean and variance for utterance duration in second are presented in this table. It is clear that there is a big variance between utterance duration for utterances used by the four segmentation method. For example, for EU1-DB the minimum duration is 1.440 second and the maximum duration is 8.978 this big difference may affect the quality of the extracted acoustic features. Therefore, segmentation into units may solve this problem and improve the quality of the extracted features for units instead of utterance.

Table 2. Utterances' duration statistics for the selected database for different segmentation methods

method	Dataset	Utterance (#)	Duration in second			
			Min	Max	Mean	Var
$EU^{(1)}$	U1-DB	200	1.440	8.978	2.819	1.513
$EU^{(2)}$	U2-DB	200	1.440	8.978	2.819	1.513
$EU^{(3)}$	U3-DB	200	1.440	8.978	2.819	1.513
$EU^{(3)}$	U4-DB	198	1.440	8.978	2.829	1.518

Table 3, presents the number of units after applying the four segmentation methods; EU1-DB, EU2-DB, EU3-DB, and EU4-DB for utterances presented in Table 1. For example, the first row in table 3 includes the number of segmented units using segmentation method EU1-DB, the 50 neutral utterance is segmented into 317 while the 50 the Happy utterances are segmented to 357 units.

Table 3. Unit distribution of emotion classes for the selected database for different segmentation methods

method	Dataset	Unit (#)	Emotion Classes			
			Neutral	Happiness	Anger	Sadness
$EU^{(1)}$	EU1-DB	1,493	317	357	328	491
$EU^{(2)}$	EU2-DB	1,293	267	307	278	441
$EU^{(3)}$	EU3-DB	1,093	217	257	228	391
$EU^{(3)}$	EU4-DB	893	167	207	178	341

Finally, Table 4, shows the duration statistics for the units presented in Table 3. The minimum, maximum, mean and variance for unit duration in second are presented in this table. It is clear that the variance between unit duration is very small for the for segmentation methods. The maximum variance is 0.10 second. Thus, the units obtained by segmentation method are standard units that have similar duration. Therefore, the extracted features for units is more reliable than those of utterance.

Table 4. Units' duration statistics for the selected database for different segmentation methods

Segmentation method	Dataset	Unit (#)	Duration in second			
			Min	Max	Mean	Var
$EU^{(1)}$	EU1-DB	1,493	0.02	1.37	0.22	0.03
$EU^{(2)}$	EU2-DB	1,293	0.06	1.65	0.43	0.06
$EU^{(3)}$	EU3-DB	1,093	0.15	1.77	0.62	0.07
$EU^{(3)}$	EU4-DB	893	0.20	2.01	0.82	0.10

4.3 Features extraction

The Mel frequency cepstral coefficients (MFCC) is extensively used for SER [17]. MFCC have good performance in description of the human ear's auditory characteristics. Therefore, the LLD acoustic features in terms of MFCC are extracted from each frame in each unit. Then applying some functional to extract the global features form each unit. For each frame, the FFT and the power spectrum were calculated. The Mel-frequency spectrum was computed. Therefore, the spectrum is filtered to mimic the human ear. Finally, the discrete cosine transform (DCT) of the Mel log powers was calculated and the first 13-order of the MFCC coefficients are extracted as shown in Figure 3. For each MFCC coefficient, 13 functional were computed (minimum, maximum, range, mean median, mode, standard deviation, variance, skewness, kurtosis, quantiles, percentiles and interquartile range) over all frames of emotion unit. Each unit's MFCCs feature is composed of a 169 points vector.

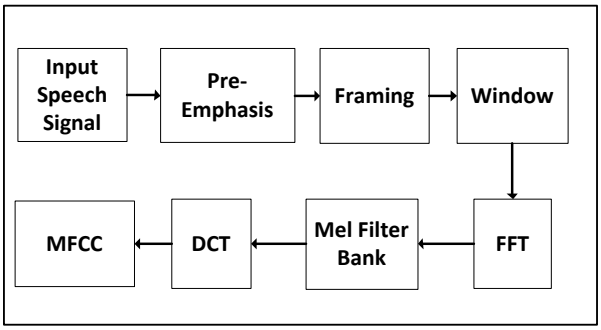

Figure 3: The Block Diagram for extracting Mel Frequency Cepstral Coefficients (MFCC)

4.4 Feature selection and classifier

To find the optimal unit, the impact of each emotion unit on SER is investigated, Thus, SVM classifier was used for evaluation. This classifier is based on statistical learning theory. Given training data: $D = \{\{x_i, y_i\},\ i = 1{:}N\}$, where $x_i \in R^d$ is the sample features represented in the d-dimensional space, $y_i \in$ {-1, +1} is the class labels, and N the number of samples. Let a linear classifier that characterized by the set of pairs (w, b) that satisfies the following inequalities of optimum hyperplane for any pattern x_i in the training set:

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

Where w is weight vector, b is the value of the tendency and ξ_i is positive artificial variable. The classification of the SVM depend on the artificial variable ξ_i as follows: if ξi=0 then the sample xi is correctly classified. If the ξi is in range; 0< ξi <1, then xi is also correctly classified however its position is among extreme planes. And when ξi >1, it is wrongly classified.

If the two-class problem cannot be linearly separated, then the kernel function is used for classification in a higher dimension. The standard kernel functions are "linear", "polynomial', "radial basis" and "sigmoid" function. SVM is basically designed for two class problems. In addition, for multiple classification, One-Against-Rest, One-Against-One, and Multi-Class Ranking approaches can be used. In the study, the kernel function RBF function was used in SVM classifier. In this study the sample x_i represents the acoustic features that were extracted in section 4.3. Since the number of features is 169, a feature selection method based on sequential forward floating search (SFFS) and SVM is applied and the best set of acoustic feature are selected. The radial basis kernel function is preferred. All values are normalized between 0-1 to remove measurement unit differences between the obtained feature sets.

4.5 Results for emotion classification

The proposed system for SER is used for evaluating which emotion unit is the optimal among the four types of units $(EU^{(1)}, EU^{(2)}, EU^{(3)}, EU^{(4)})$. These types are used for segmenting EMO-DB separately. Four datasets of emotional units are obtained, namely, EU1-DB, EU2-DB, EU3-DB, and EU4-DB. To measure the impact of each unit on the recognition rate of emotional state, the four datasets were used separately to train and test the proposed SER system using 5-fold cross validation. The inputs to this system are acoustic features for each dataset and the output is emotional class. The classification rate for all datasets are presented in Figure 4.

It clear from this figure that the emotion unit $EU^{(4)}$ method attained the highest recognition rate. Therefore, this method for segmentation is considered the optimal unit. The finding of this investigation is as follow, to analyze the speech in continuous tracing for detecting the emotional changes, it is required to segment the utterance into units that has a windows of four consequence voiced segments with overlap of three voiced segments. Moreover, in order to compare with the previous study, the label of whole utterance is predicted from the predicted labels of its units using the majority voting. Therefore, the proposed system not only tracking the change of emotional state during utterance but also can predict the overall emotional state in the whole utterance.

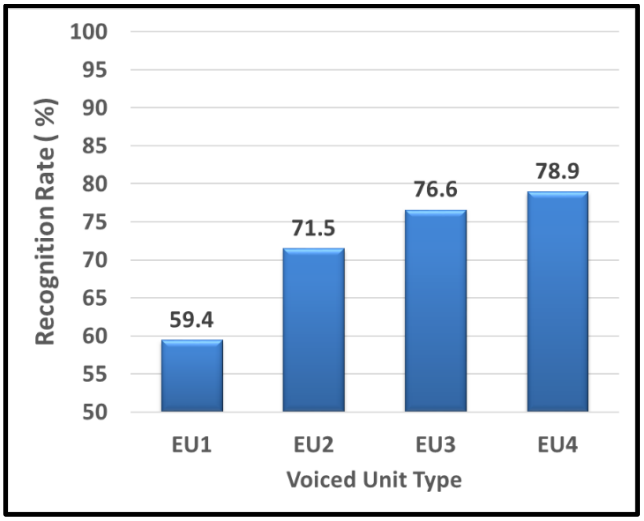

Figure 4: Recognition rate for continuous SER using the four candidates for emotion unit

The results for emotion classification for the whole utterance using the proposed method and the conventional method are presented in Figure 5.

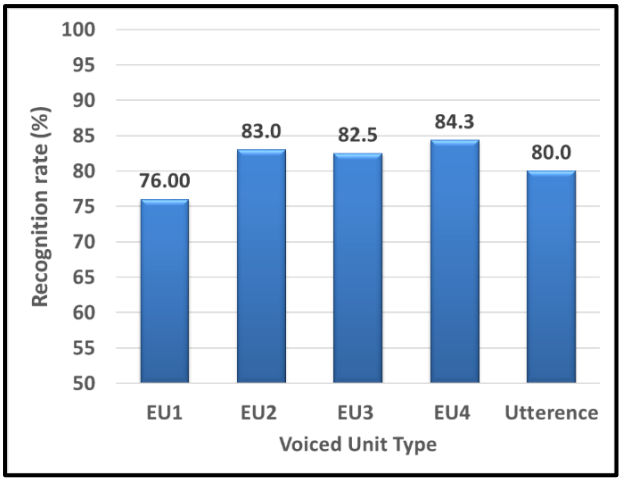

Figure 5: Utterance classification rate using the four candidates for emotion unit and traditional method

The classification results using the proposed methods outperforms the conventional method that use the utterance as unit for recognition. This result supports our assumption since the recognition rate increased by 4.3 % using the proposed method.

The obtained results were compared with previous studies that uses the same corpus to emphasis the effectiveness of the proposed method. The proposed method based on voiced segments is compared with the traditional methods that uses the whole utterance as one unit. It is found that the proposed method outperforms the three-layer method for emotion classification proposed by Elbarougy et. al. in (2014) [22]. The improvement of the proposed method is 9.3% from 75% to 84.3%. Moreover, our results for emotion classification outperform the obtained by PSO-FIS by 4.3% [23]. In addition, the proposed method outperforms the traditional method even with applying feature selection as presented in [24] the improvement is from 82% to 84.3% i.e. 2.3%. From this discussion it is clear that the proposed method for emotion classification has the ability to accurately predict the emotion category using the voiced segments, with a very high emotion classification rate.

5. CONCLUSIONS

The aim of the paper is to find an appropriate segmentation unit to be used for segmenting a speech utterance into units that are representative for emotions. Four candidates for emotion unit were proposed as follows, $EU^{(1)}, EU^{(2)}, EU^{(3)}, EU^{(4)}$. The definition of each unit is based on the number of voiced segments. The impact of each unit on SER for is evaluated using SVM classifier. To train the classifier the 13-order of the MFCC coefficients are extracted, and 13 functional were computed for each coefficients. Each unit's MFCCs feature is composed of a 169 points vector. Then a feature selection method was applied for features of each unit separately. The experimental results reveal that the $EU^{(4)}$ attained the highest recognition rate.

To compare the proposed method with the conventional method that uses the utterance as unit for recognizing the emotional state of the whole utterance. Thus the predicted labels of units were used to predict the emotional state of the whole utterance using majority voting. The classification results using the proposed methods outperforms the conventional method, the improvement in recognition accuracy was from 80% to 84.3%.

In future the proposed method for tracking the continuous emotional state will be evaluated for emotion dimension representation. Moreover, validating performance of the continuous SER using different acoustic features.

6. REFERENCES

[1] Jiang W., Wang Z., Jin J.S., Han X., Li C. Speech Emotion Recognition with Heterogeneous Feature Unification of Deep Neural Network. *Sensors (Basel)*. 2019 Jun 18;19(12):2730.

[2] Alonso-Martín, F., Malfaz, M., Sequeira, J., Gorostiza, J. F., and Salichs, M. A., "A multimodal emotion detection system during human–robot interaction," Sensors, vol. 13, no. 11, pp. 15 549–15 581, 2013.

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

[4] Vogt, T., Andr´e, E., "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," In: Proceedings of International Conference on Multimedia & Expo., Amsterdam, The Netherlands (2005).

[5] Vogt, T., Andr´e, E., and Wagner, J., "Automatic recognition of emotions from speech: a review of the literature and recommendations for practical realization," In: Peter, C., Beale, R. (eds.) Affect and Emotion in Human-Computer Interaction. LNCS, vol. 4868. Springer, Heidelberg (2007)

[6] Kerkeni, L., Serrestou, Y., Raoof, K., Cléder, C., Mahjoub, M., Mbarki, M. (2019). "Automatic Speech Emotion Recognition Using Machine Learning"

[7] Batliner, A., Fischer, K., Huber, R., Spilker, J., Noth, E.: "How to find trouble in communication," Speech Communication 40, 117–143 (2003)

[8] Batliner, A., Seppi, D., Steidl, S., and Schuller, B., "Segmenting into adequate units for automatic recognition of emotion-related episodes: a speech-based approach," Advances in Human-Computer Interaction, 2010. vol. 2010, Article ID 782802, 15 pages.

[9] Seppi, D., Batliner, A., Steidl, S., Schuller, B., and Noth, E., "Word Accent and Emotion," In Proc. Speech Prosody 2010, Chicago, IL, 2010.

[10] Burkhardt, F., Paeschke, A., Rolfes, M.; Sendlmeier, W., and Weiss, B., "A Database of German Emotional Speech," INTERSPEECH (2005).

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

[12] Ringeval, F., & Chetouani, M. (2008). "A vowel based approach for acted emotion recognition," In INTERSPEECH 2008 9th annual conference of the international speech communication association, Brisbane, Australia, 22–26 September.

[13] Deb S., and Dandapat, S., "Emotion classification using segmentation of vowel-like and non-vowel-like regions," IEEE Transactions on Affective Computing, 2017

[14] Moattar, M., and Homayounpour, M., "A simple but efficient real-time voice activity detection algorithm," in EUSIPCO. EURASIP, 2009, pp. 2549–2553.

[15] Moattar, M., Homayounpour, M., and Kalantari, N., "A New Approach for Robust Real-time Voice Activity Detection Using Spectral Pattern," ICASSP, 2010, pp. 4478-4481.

[16] H. Kawahara, and I.M.-katsuse, and A.D. Cheveign, "Restructuring speech representations using a pitch adaptive time-frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds," Speech Communication, vol. 27, pp. 187–207, 1999.

[17] El Ayadi, M., Kamel, M. S., Karray, F., "Survey on speech emotion recognition: Features, classification schemes, and databases," Pattern Recognition 44 (3) (2011) 572{587.

[18] Busso, C., Lee, S., and Narayanan, S., "Analysis of emotionally salient aspects of fundamental frequency for emotion detection," IEEE Trans. Audio, Speech Language Process., vol. 17, no. 4, pp. 582–596, May 2009.

[19] Lee, C., M., Yildirim, S., Bulut, M., Kazemzadeh, A., Busso, C., Deng, Z., Lee, S., Narayanan, S., "Emotion recognition based on phoneme classes,", in Proc. of ICSL, 2004.

[20] Eyben, F., Ẅollmer, M., Graves, A., Schuller, B., Douglas-Cowie, E., and Cowie. R., "On-line emotion recognition in a 3-d activation-valence-time continuum using acoustic and linguistic cues," Journal on Multimodal User Interfaces, 3(1-2):7–19, Mar. 2010.

[21] Mansoorizadeh M., Charkari N.M., (2007) "Speech emotion recognition: comparison of speech segmentation approaches," In: Proceedings of IKT, Mashad, Iran

[22] Elbarougy, R. and Akagi, M. "Improving Speech Emotion Dimensions Estimation Using a Three-Layer Model for Human Perception," Journal of Acoustical Science and Technology, 35, 2, 86-98, March, 2014.

[23] Elbarougy, R. and Akagi, M., "Optimizing fuzzy inference systems for improving speech emotion recognition," Advances in Intelligent Systems and Computing, vol. 533, pp. 85-95, 2017.

[24] R. Elbarougy, M. Akagi, "Feature selection method for real-time speech emotion recognition," in Co-ordination and Standardization of Speech Databases and Assessment Techniques (CO-COSDA), 2017 20th Oriental Chapter of the International Committee for the. IEEE, 2017, pp. 1–6.