Continuous Tracking of Emotional State from Speech Based on Emotion Unit

Reda Elbarougy [1;2], *B.T. Atmaja* [1] *and Masato Akagi* [1]
(1) Japan Advanced Institute of Science and Technology, Japan
(2) Dep. of Math., Faculty of Science, Damietta University, Egypt

1 Introduction

Continuous tracking of emotional state from speech is a challenging problem. Due to the essential need for audio segmentation into an appropriate unit for emotions [1]. This unit is called emotion unit. Previous studies of acted emotions have been based on the whole utterance as an emotion unit for recognizing the emotional state. It is assumed that the emotional state is fixed during one utterance. However, in continuous speech this kind of observable unit does not exist. Neither is the segmentation into utterances straight-forward nor can a constant emotion be expected over an utterance. Thus, considering utterance as an emotion unit is not suitable for speech emotion recognition (SER) especially for long utterance. Therefore, a new emotion unit based on voiced segments is proposed for continuous tracking of emotional state.

Linguistically motivated emotion units such as utterance has many limitations for continuous speech. It requires preprocessing by an automatic speech recognition (ASR) system to obtain unit boundaries. Time constraints make it mandatory not to wait with ASR and other processing modules until the speaker has finished his/her full turn. Emotion is dynamic and may changes during the utterance, especially in spontaneous speech, utterances' duration is much varied from very short to very long utterance. Therefore, long utterances may include different emotional states. Thus, the extracted low-level descriptors (LLDs) of acoustic features such as Mel-frequency cepstral coefficient (MFCC) from such utterances are inconstant because they representing different emotional states. As a result, applying functional statistics such as (mean, standard deviation) to obtain global statistic form the LLDs for long utterance are not reliable. To emphasis the discriminative properties of acoustic features over one unit, it is needed to find a standard emotion unit to obtain reliable statistics.

The goal of this study is to find an appropriate emotion unit that can be used for analysis, extraction of emotion-relevant features, and classification. Moreover, to investigate the feasibility of this unit for continuous tracking of emotional state [1]. Finding the appropriate emotion unit that include only one emotional state, make the extracted LLDs features from this unit more consistent. Thus, applying some functional to extract the global feature leads to more expressive features. Segmentation of an utterance into its emotion units may help in accurately determining the emotional state of each unit. As a result, we can determine the ongoing emotional state in real-time in case of continuous speech. The proposed emotion unit based on voiced segments is described in details in section 3. To evaluate this method, SER system based on the dimensional approach using support vector machine is used. For validating it, the emotional database EMO-DB is used.

2 Speech material

The Berlin emotional speech database (EMO-DB) is chosen for evaluating the proposed method [2]. There are ten utterances from ten different actors, five males and five females. These ten utterances are divided into five short sentences (1.5s approximately) and five longer sentences (4s approximately). These sentences were not equally distributed between the various emotional states. For training purpose, an equal distribution of the four emotional states was used, 50 happy, 50 angry, 50 sad, and 50 neutral; in total, 200 utterances were selected. EMO-DB originally annotated using categorical representation. To evaluate the proposed method based on dimensional representation, it is required to re-annotate it using this representation. The emotional state is represented as a point in a two dimensional space, i.e. valence-arousal space. Thus, using a listening test each utterance in selected dataset was labeled in terms of valence and arousal using a 5-point scale {-2, -1, 0, 1, 2}. Valence scale is very negative (-2), negative (-1), neutral (0), positive (1), and very positive (2). Arousal scale is very calm (-2), calm (-1), neutral (0), excited (1), and very excited (2).

3 Proposed emotion unit

In this study, it is assumed that emotion unit should be investigated within the voiced segments. These segments contain F0 information that are mostly used to express the emotional state of the speaker. Voiced segments of an utterance include vowels which are very important for SER, because vowels are the rich part with emotional information [3].

Segmentation of speech utterance into its vowels is very challenging task and require either prior knowledge such as the phoneme boundaries or using an ASR system to find these boundaries. On the other hand, segmenting into voiced segments can be easily done using voice activity detection (VAD) with a very high performance. 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 can be done by using STRAIGHT software [4]. The algorithm used for this segmentation is based only on acoustic information that make it easy to be implemented in real-time. Suppose the utterance U_i is segmented into its voiced segments using this algorithm, the output of segmentation are the waveforms of all voiced segments which given by

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

where i is utterance index, $V(U_i)$ 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 U_i. Figure 1 shows the segmentation of U_i into its voiced segments.

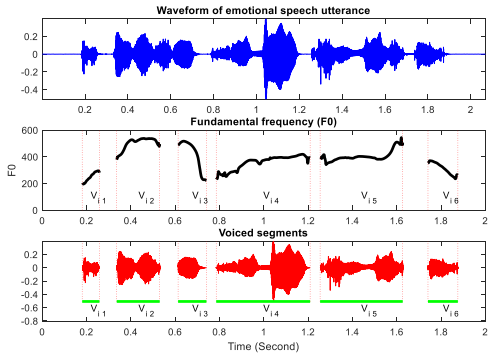

Figure 1: Segmentation of speech utterance U_i into voiced segments using VAD, i utterance index.

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. In this study, it is assumed that one voiced segment cannot include more than one emotion i.e. during one voiced 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 of voiced segments. 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. Thus, we define emotion unit in terms of number of voiced segments. For example, emotion unit 1 ($EU^{(1)}$) is the method that segments an utterance into units/segments include one voiced segment as given by

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

where V_{ij} , M_i are as explained in equation (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 emotion unit 2 ($EU^{(2)}$) that segments utterance into units that contains two consequence voiced segments in each unit as given by

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

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.

In general, the definition of emotion unit k ($EU^{(k)}$) by

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

$EU^{(k)}$ segments the utterance U_i into units which consists of k voiced segments. From the above definition, it is clear that, number of emotion units in one utterance depends on both the number of voiced segments and the type of unit representation.

4 Evaluation of proposed emotion unit

Given an emotional speech database $DB = \{U_i, i = 1:N\}$ of N emotional utterances. Appling different emotion unit segmentation on DB, different datasets of emotion units are obtained. For example, applying $EU^{(k)}$ that include k voiced segments we obtain the following dataset

$$EU^{(k)}(DB) = \left\{ \begin{array}{c} S_{ij} = \cup_{l=j}^{l=j+k-1} V_{il}, \\ i = 1:N, j = 1:M_i - k + 1 \end{array} \right\} \qquad (5)$$

The obtained units using this segmentation method 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 emotion unit, the impact of including different number of voiced segments in the proposed unit on SER is investigated. The unit that yields the highest recognition accuracy for SER system is the optimal one. In this study, the investigation for emotion unit is based on the

dimensional representation of emotion. Thus, traditional problem statement for emotion dimension estimation is reformulated according the concept of emotion unit. Traditionally, SER based on the dimensional representation using utterance unit can be defined as follows: given a dataset of emotional speech utterances, each utterance is annotated using dimensional approach. The sequence of labels for emotion dimension valence and arousal is given by

$$E^{(d)} = \left\{E_i^{(d)},\ i = 1{:}N\right\} \qquad (6)$$

where $\mathrm{d} \in (\mathrm{valence}, \mathrm{arousal})$, and $E_i^{(d)}$ is the value of emotion dimension d for utterance U_{i}. The conventional task of SER is how to construct and train an SER system to estimate emotion dimensions d for a new utterance.

Using emotion unit concept, the conventional approach could be reformulated as follows, suppose for example, $EU^{(k)}$ is used for segmentation, the obtained dataset as given by equation (5). Since the label of each unit is not given in the original database, therefore, it is assumed that each emotion unit S_{ij} has the label of the utterance U_{i} which belong to. As a result, the emotion dimensions' values of all units in the obtained database are

$$X^d = \left\{ \begin{matrix} X_{ij}^{(d)} = E_i^{(d)}, \\ i = 1{:}N, \quad j = 1{:}M_{\mathrm{i}} - k + 1 \end{matrix} \right\} \quad (7)$$

where k is the number of voiced segments in the unit S_{ij}, $X_{ij}^{(d)}$ is value of emotion dimension d, $\mathrm{d} \in (\mathrm{valence}, \mathrm{arousal})$ for the unit S_{ij}.

Therefore, the new definition for emotion estimation problem is as follow; given a dataset of emotion units as defined by equation (5) and the values of emotion dimensions for all units by equation (7), how to predicted emotion dimensions values of valence and arousal for a new unit S_{ij} which the system not trained on. Since, one utterance consists of number of emotion units, therefore, predicting the emotional state of each unit is considered a continuous tracking of emotional states within one utterance.

4.1 Speech emotion recognition system

The proposed system for detecting the emotional state is composed of two stages, training stage and testing stage. In training stage, utterance is segmented into its units as explained in section 3. Then acoustic features extracted from each unit. The final step is to train the proposed estimator to learn the relationship between acoustic features extracted from units and the emotional state of these units. Moreover, in testing stage, the trained system is used to predict the emotional state of a new utterance. This stage includes 3 steps: the first step is used to segment the input utterance into its emotion units. Then extract acoustic features from each emotion unit. In addition, these features are used as inputs to the trained estimator to predict the emotional state of each emotion unit.

To avoid overfitting, the whole dataset of emotion units is divided using 5-fold cross validation. Thus finally the predicted emotion dimension of all utterances are given by

$$Y^d = \left\{ \begin{matrix} Y_{ij}^{(d)}, \\ i = 1{:}N, \quad j = 1{:}M_{\mathrm{i}} - k + 1 \end{matrix} \right\} \quad (8)$$

where $Y_{ij}^{(d)}$ is predicted value of emotion dimension d, of the unit S_{ij}.

The Mel frequency cepstral coefficients (MFCC) is widely used for SER. MFCC have good performance in description of the human ear's auditory characteristics. Therefore, the LLDs of acoustic features in terms of MFCC are extracted from each frame in each unit. For each frame, the first 13-order of the MFCC coefficients are extracted. For each MFCC coefficient, 13 statistical values 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 vector is composed of a 169 elements.

4.2 Results for emotion classification

Support vector regression (SVR) estimator was used to estimate valence and arousal from the extracted acoustic features. To measure the impact of each unit on the estimation accuracy of emotional state, the performance of the proposed system for SER is evaluated for using different emotion units. Segmenting the original database using a specific emotion unit yields a new dataset of units. The inputs to the SER system are acoustic features form the obtained dataset and the output is values of emotion dimensions.

To evaluate effectiveness of the used emotion unit segmentation method, continuous mean absolute error (CMAE) was used as given by

$$CMAE^{(d)} = \frac{\sum_{i=1}^{N} \sum_{j=1}^{M_{\mathrm{i}}} \left| Y_{ij}^{(d)} - X_{ij}^{(d)} \right|}{\sum_{i=1}^{N} M_{\mathrm{i}}} \qquad (9)$$

where $\mathrm{d} \in \{\mathrm{Valence}, \mathrm{Arousal}\}$ and $Y_{ij}^{(d)}$ is the

output of the proposed system, $X_{ij}^{(d)}$ is the reference value of emotion dimensions, N is the number of utterances in used emotion corpus.

It is important to note that, when using $EU^{(k)}$ short utterances that contains less than k voiced segments must be exclude from analysis. To keep as much as possible from the original utterances, only the first four emotion units were used. Prediction performance for emotion units that include one, two, three and four voiced segments shown in Figure 2.

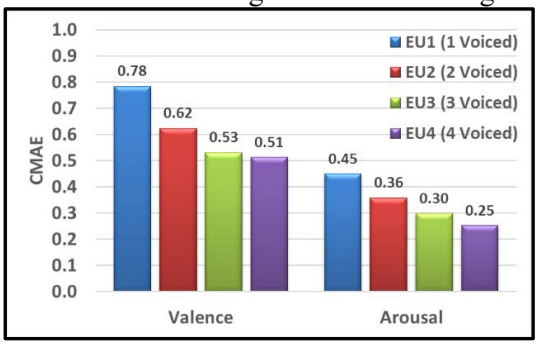

Figure 2: Prediction performance for continuous SER using the four candidates for emotion unit

It clear from this figure that $EU^{(4)}$ method attained the highest recognition rate. Therefore, this method for segmentation is considered the optimal unit. The finding of this study is that, to continuously tracing 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, the predicted values of emotion dimensions for all units can be used to determine the overall emotional state of the whole utterance by using the mean value for emotional state of units using the following equation.

$$Y_i^{(d)} = \frac{\sum_{j=1}^{M_i} Y_{ij}^{(d)}}{M_i} \qquad (10)$$

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. The mean absolute error (MAE) is used to measure the performance of the estimation accuracy for the whole utterance as given by

$$MAE^{(d)} = \frac{\sum_{i=1}^{N} \left| Y_i^{(d)} - E_i^{(d)} \right|}{N} \qquad (11)$$

The result of the SER is presented in Figure 3. It is clear that the prediction performance using the proposed methods outperforms the conventional method that uses the utterance as unit for recognition. This result supports our assumption since both emotion dimensions are improved using the proposed method of emotion unit. The improvement for valence was from 0.68 to 0.52, and the improvement for arousal was from 0.34 to 0.21.

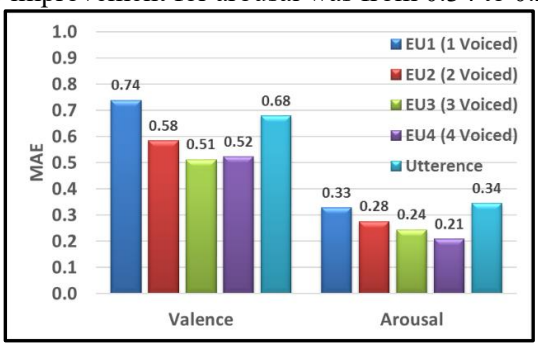

Figure 3: Prediction performance using the four candidates for emotion unit and traditional method

5 Conclusions

The aim of this paper is to continuously tracking the emotional state in terms of predicting valence and arousal values. To accomplish this task, a segmentation method based on emotion unit is proposed. This unit defined in terms of voiced segments of each utterance. To find the optimal unit, the impact of number of voiced segment in emotion unit on the prediction performance is investigated. The unit that attain the highest performance is the optimal one. The experimental results reveal that the $EU^{(4)}$ attained the highest performance in terms of mean absolute error. The predicted values of emotion dimensions for each unit were used to predict the emotional state of the whole utterance using mean value of emotion dimensions. The prediction performance for using the proposed methods outperforms the conventional method for both valence and arousal.

6 Acknowledgment

This study was supported by Japan Society for the Promotion of Science (JSPS).

7 References

[1] T. Vogt. “Real-time automatic emotion recognition from speech.” PhD thesis, Technischen Fakultät der Universität Bielefeld, 2010.

[2] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier, B. Weiss, “A Database of German Emotional Speech,” INTERSPEECH (2005).

[3] Ringeval, F., & Chetouani, M. “A vowel based approach for acted emotion recognition,” In INTERSPEECH (2008), Brisbane, Australia, 22–26 September.

[4] 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.