FEATURE SELECTION METHOD FOR REAL-TIME SPEECH EMOTION RECOGNITION

Reda Elbarougy [1;2], *and* *Masato Akagi* [1]

[1] Japan Advanced Institute of Science and Technology, Japan
[2] Dep. of Math., Faculty of Science, Damietta University, Egypt

elbarogy2000@yahoo.com, akagi@jaist.ac.jp

ABSTRACT

Feature selection is very important step to improve the accuracy of speech emotion recognition for many applications such as speech-to-speech translation system. Thousands of features can be extracted from speech signal however which group of features are the most related for emotion states of the speakers. Until now most of related features to emotional states are not yet found. The purpose of this paper is to propose a feature selection method which have the ability to find most related features with linear or non-linear relationship with the emotional state. Most of the previous studies use either correlation between acoustic features and emotions as for feature selection or principal component analysis (PCA) as a feature reduction method. These traditional methods does not reflect all types of relations between acoustic features and emotional state. They only can find the features which have a linear relationship. However, the relationship between any two variables can be linear, non-linear or fuzzy. Therefore, the feature selection method should consider these kind of relationship between acoustic features and emotional state. Therefore, a feature selection method based on fuzzy inference system (FIS) was proposed. The proposed method can find all features which have any kind of above mentioned relationships. Then A FIS was used to estimate emotion dimensions valence and activations. Third FIS was used to map the values of estimated valence and activation to emotional category. The experimental results reveal that the proposed features selection method outperforms the traditional methods.

Index Terms— Acoustic features, Feature selection, Feature reduction, Speech emotion recognition, Emotion dimensions, Real-time recognition, Fuzzy inference system.

1. INTRODUCTION

Speech-To-Speech Translation (S2ST) is the process in which the input speech is translated from one language to another language [1, 2]. The output of S2ST system is usually neutral speech, however in order to make it more natural speech it is should be colored with emotion [3]. Therefore it is important to design and develop an accurate emotion recognition system. For robust speech emotion recognition, it is indispensable to select the most effective and significant features [4]. These selected features is crucial to accurately predicting the emotional state. The speech emotion recognition (SER) system should be implemented on the real-time due to it needed to be embedded on a S2ST system. Therefore, it is also required for the extracted acoustic features to be easily calculated in a real-time. The purpose of this paper is to introduce a feature selection method to be used for predicting the emotional state in a real time SER system.

Still speech emotion recognition is a challenging problem for many reasons [5]: It is not known what are the most important segments of an utterance for emotion recognition? It is also not known which features is essential for predicting the emotional state? [6, 7] Most of the previous studies use the utterance level i.e. for calculating the acoustic features by using the average of the specified acoustic features for all frames of the speech signal [8]. However, the utterance may include vowels and consonants. Therefore, it is important to add features calculated in smaller level such as vowel or consonant to the utterance level [7]. Previous studies focused only on using acoustic features extracted on the utterance level. Spectral feature are more stable for vowel segments. Some features have no values in the consonant segment. Therefore more effective features will be included in the vowel segment. However vowel level is ignored in many of them. This study focused on vowel part for many reasons. Boundary of vowel segment can be easy to be determined automatically using voice activity detection method. Vowels are more riches by emotional characteristics.

Emotional state can be represented by two methods categorical and dimensional method [9]. The first method is represents emotion by several classes such as happy, angry, sad and neutral. The second method representing emotional state as a point on 2-dimenional space called valence-activation space. It is also found that the recognition accuracy is optimized by using a hybrid method which combine both representations. Such as the method used in [7] by Grimm et al. in their study by firstly estimating emotion dimensions valence and activation followed by recognizing the emotional

state by using the estimated dimensions. However, in order to improve the recognition accuracy the estimation of the two dimensions must be improved. Therefore, it is important to find the related acoustic features for emotion dimensions. Thus it is important to answer this question what are acoustic features relevant to emotion dimensions?

Although the valence dimension was found to be more difficult to be estimated due to lack of related acoustic features [10, 11]. To find the most related acoustic features several studies used the correlation between acoustic features and emotion dimensions [6, 8, and 10]. However the correlation between acoustic features and emotion dimension is not enough with the most difficult dimension valence. Since the correlation between two variables can describe the linear relationship between them but cannot find if there is a non-linear relationship between them. On the other hand, for the activation dimension from most of the previous study was very easy to be estimated [10]. Correlation could be a good for linear relationship between AF and emotion dimension, such as the relation between activation and acoustic features. The current study believe that the relation between acoustic features and activation is linear. Therefore the correlation can find the most related acoustic features. Thus many studies claim that they have found many acoustic features related to this dimensions.

Feature selection is refers to the process of reducing the features for processing and analysis, or of finding the most meaningful features. This process speeds up learning model, and also improves the accuracy [ref]. Feature reduction is means reducing the number of dimensions of the feature. Dimensionality reduction is useful to speed up algorithm execution, and help with the final classification/clustering accuracy as well. Principal Component Analysis (PCA) is widely used tool for dimensionality reduction of features, providing a linear combination of all features that maximizes data variance. PCA is a technique that seeks a reduced-dimensional basis that best captures the variance of the data. The direction with the largest projected variance is called the first principal component. The orthogonal direction that captures the second largest projected variance is called the second principal component [ref]. However, PCA does not depend on the label or values of emotion therefore the results may not optimal. Therefore in this study we propose a feature selection method which take into account the label or values of emotion to select the acoustic feature related to the desired emotion dimension. This method deal with unlabeled data and tries to maximize variance it is applied on acoustic features only without any information about label valence or activation i.e. it does not depend on output of the required emotion recognition system.

From the above introduction since the relation between any two variables can be linear, non-linear of fuzzy. Acoustic feature selection method must take into consideration the above types of relations. Therefore, a feature selection method based fuzzy inference system was proposed. The proposed feature section method can find all these kinds of relations between acoustic features (AF) and emotion dimension (ED).

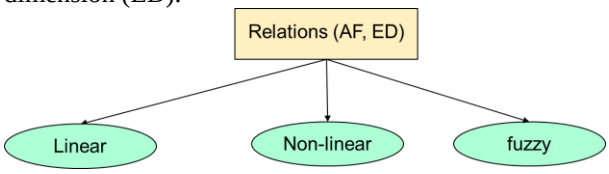


Fig. 1. Types of relations between acoustic features (AF) and emotion dimensions (ED).

The proposed feature selection method can not only find the linear relations but also non-linear relations. The proposed method is used to overcome the drawback of the traditional approach for feature selection. Such as the limitation of correlation for selecting acoustic features for valence dimension.

2. PROPOSED STRATGY

In order to realize affective S2ST system it should be on-line or real-time system to be embedded on a variety devices to convey the speaker emotional state in the output language. However, the challenge for constructing this system is how to recognize the source emotional state of the target speech on real-time? Online system require an automatic extraction of acoustic features without any prior knowledge. Therefore this study tries to answer what are the important segments of a speech utterance? i.e. segments which carry more emotional information? And how to automatically extract acoustic features from these segments? This is still a challenging problem.

In order to construct real-time affective S2ST system it is important to investigate the followings:

1. How important of feature selection method for SER for our proposed S2ST system?
2. How to select the related feature for the challenging emotion dimension valence?
3. What are most important segments of an utterance for emotion? Is it vowel segment, consonant segment or the whole utterance?
4. Whether SER system can be applied for different languages? Source language can be any language.
5. How to automatically extract acoustic feature from the proposed segments without any prior knowledge?

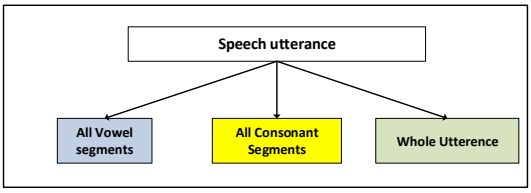


Fig. 2. Segments of speech utterance

Previous studies focused only on using acoustic features extracted on the utterance level [12]. Spectral feature are steadier for vowel segments. Therefore more effective features will be included in the vowel segment. However, vowel level was ignored in many previous studies. This study focused on vowel part for many reasons. Boundary of vowels is easy to be determined automatically using voice activity detection (VAD). Vowels are more riches by emotional characteristics. In order to answer this question: what are acoustic features (AF) relevant to emotion dimensions?

Many feature selection methods were proposed however several of them were based on the correlation between acoustic features and emotion [8, 10]. Correlations method can be good for linear relationship between AF and emotion dimension, such as the relation between activation and acoustic features [13]. However, the correlation based feature selection method has a limitation for selecting acoustic features for valence [14]. In order to avoid this drawback the feature selection method must take into consideration many types of relations. The relation between two variables could be linear, non-linear or fuzzy relationship. Therefore, a feature selection method based on fuzzy inference system FIS was proposed. All acoustic feature values will be extracted automatically by firstly applying VAD to find the vowel boundaries and then applying the features extraction for the selected acoustic features. The extracted features will be used as an inputs to the affective S2ST system.

3. FEATURE EXTRACTION

The first step of constructing speech emotion recognition system is the feature extraction step. In this step thousands of features could be extracted. The features will be the input of the proposed SER system. In order to extend the feature domain as explained in the previous section two additional ideas will be developed. The first is use the vowel, consonant as well as the utterance level. The second is using different statistics such as; minimum, maximum, mean and range. In this study the traditional MFCC coefficient were extracted for each frame. The frames are divided into consonant, vowel frames. The above statistics were applied to the 13 MFCC coefficients for the frames of vowel and consonant segment individually as well as applied for the frames of the whole utterance as shown in the figure 3. Finally the extracted feature vector contains 156 features. For each segment, there are 13 MFCC coefficients and 4 statistics. Therefore 52 features for each segment multiplied by 3 segments which yield 156 features.

However, for the reason that vowel are more stable for cepstral features, the vowel level and the utterance levels were used. The proposed vowel-level for acoustic features extraction can be used for automatic emotion recognition. It is also possible to extend the number of features using several statistic for each segment. These feature will be investigated in our future work.

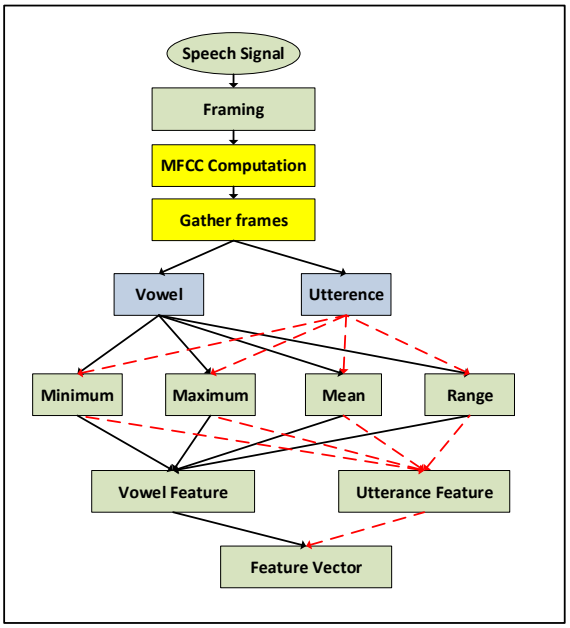


Fig. 3. The extracted acoustic features and the used segments and statistics.

4. FEATURE SELECTION

It is not clear which features that extracted in the previous section is more important for emotion recognition. Therefore a feature selection method is needed for this purpose. This section explains the proposed feature selection method. This method can find all kinds or relationship between acoustic features and emotion dimensions specifically the non-linear and fuzzy relation which are very complicated to be found by the traditional correlation method.

Fuzzy inference system (FIS) is usually used as mathematical tool for approximating non-linear functions. This model can import qualitative aspects of human knowledge and reasoning process by data sets without employing precise quantitative analysis. Most of the statistical methodology such as the correlations mainly based on a linear and precise relationships between the input and the output, while the relationship between acoustic features and emotion dimensions are non-linear. Therefore, fuzzy logic is a more appropriate mathematical tool for describing this non-linear relationship.

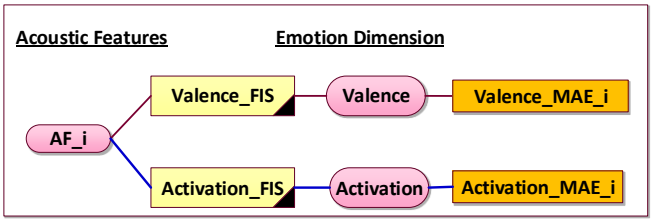


Fig. 4. Using FIS for acoustic feature selection

The reasons are as follows: (1) Fuzzy logic is a tool for embedding existing structured of human knowledge into

mathematical models [15] using If-Then rules, and this is exactly what the model proposes to do in dealing with the perception of expressive speech. (2) Fuzzy logic models non-linear functions of arbitrary complexity [16], and the relationship between emotion dimensions and acoustic features are certainly complex and non-linear. Therefore, fuzzy logic is appropriate to model these relationships.

The proposed feature selection method is presented in the following figure. For any speech database the desired acoustic features can be extracted as explained in the previous section. On the other hand emotion dimensions valence and activation were estimated using listening test using human subjects. In order to select the best AF related with emotion dimensions one FIS is used to model the relation between AF_i and ED. This can be done by dividing the data into two sets training and testing sets using. The first set is used to train the FIS model in order to mimic the relation between the AF_i and ED. The second set used to test the trained model. The Mean Absolute Error (MAE) is used as an evaluation criteria, the smallest MAE the best estimation of ED from AF. Threshold is used to specify if the acoustic feature could be selected or not as shown in Fig 5.

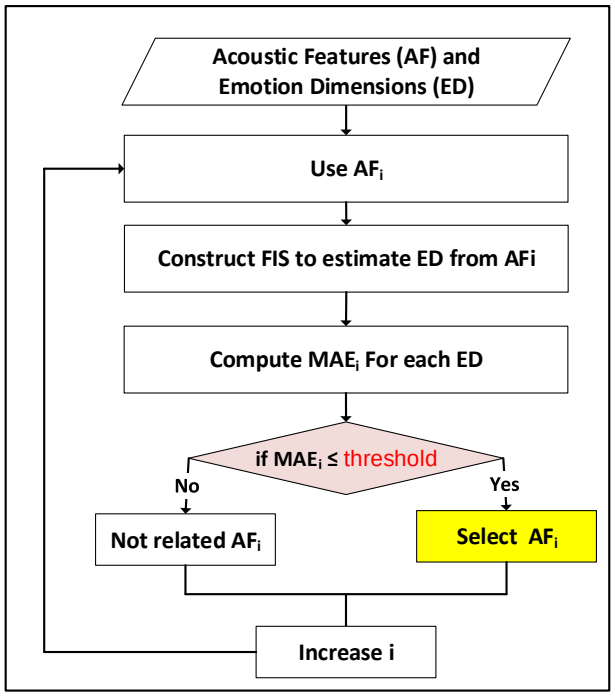


Fig. 5. The proposed feature selection method.

Correlation between acoustic feature and emotion dimension may be very weak. However, there is a non-linear relation between them as shown in the following table. This table presents the top 10 acoustic features which gives the smallest mean absolute error (MAE). This table also shows the correlation between acoustic features and emotion dimensions. The experimental results for the Japanese database which was used previously in our study [14]

The following table shows the results of implementing the proposed feature selection method. The top 10 acoustic features strongly related with valence dimension are listed in a descending order of MAE. The first column list the name of the used acoustic feature. The second column present the correlation between acoustic feature in the first column and the valence dimension. The third column gives the MAE for estimating the valence dimension from the corresponding acoustic future.

Table 1: the selected acoustic features for valence, results sorted in descending order for MAE.

AF Name	Corr(AF,ED)	MAE_i
Min_Utter_MFCC_0	-0.15	0.22
Mean_V_MFCC_1	-0.10	0.44
Mean_Utter _MFCC_1	-0.11	0.46
Range_Utter _MFCC_0	0.23	0.52
Mean_V_MFCC_1	-0.17	0.55
Mean_V_MFCC_10	0.49	0.57
Min_V_MFCC_0	-0.05	0.58
Max_V_MFCC_10	0.47	0.59
Mean_Utter_MFCC_10	0.47	0.59
Mean_Utter_MFCC_4	-0.63	0.60

From this table it is clear that even the acoustic feature (Min_Utter_MFCC_0) in the first row have a very low correlation although it have the smallest and best MAE. Using the correlation method this acoustic feature certainly will be excluded from the selected features however it is the best feature.

5. PROPOSED SER SYSTEM

In order to investigate the effectiveness of the proposed feature selection method a SER system was implemented. The proposed system consists of two stages in the first stage the values of valence and activation was estimated from the input acoustic features. In order to take the advantage of improving emotional classification by mapping emotion dimension to categories. Another stage is required to map the estimated emotion dimensions to the emotion category. In the first stage two FISs were used to estimate emotion dimensions from the selected features in the previous section. Then a third FIS was used to implement the second stage as shown in Fig 3.

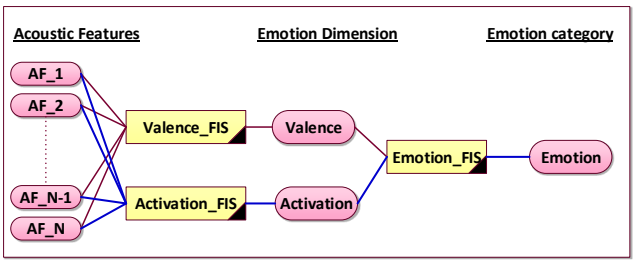


Fig. 6. The proposed SER system for emotion recognition

6. RESULTS

Firstly, in order to answer whether the vowel level can give us more valuable features and can improve the recognition accuracy or not. Therefore, 10 features were selected from vowel level and utterance level. Moreover, from both level combined together. The results of the recognition accuracy were presented in the following tables.

Table 2: Japanese database: The recognition accuracy using different level.

Level	Mean Absolute Error (MAE)		%
	Valence	Activation	Recognition Rate
Utter	0.53	0.21	86.6%
Vowel	0.34	0.31	90.5%
Both	0.22	0.18	97.8%

Table 3: German database: The recognition accuracy using different level.

Level	Mean Absolute Error (MAE)		%
	Valence	Activation	Recognition Rate
Utter	0.99	0.35	61.0%
Vowel	0.70	0.35	73.5%
Both	0.70	0.34	74.0%

From these tables it clear that the utterance level << vowel level << both levels. Therefore, in order to find a strong acoustic features, it is important to investigate the vowel segment. Including vowel segment will improve the recognition accuracy as shown from the above tables.

In order to investigate the effectiveness of the proposed feature section method, it is required to compare the proposed method with the traditional methods. Therefore, the proposed SER was trained using three groups of features. In this study, 10 features were selected for each group. The first group of features were selected using the correlation method. The second group is obtained by reducing the number of features to 10 features using the Principal Component Analysis (PCA). The third group is obtained by selecting the 10 features using the proposed method explained in section 4.

The second and most important issue for this paper is to compare our proposed feature selection method with the traditional one. Using the 10 features selected from both levels vowel-utterance. The SER system was implemented using these 10 selected acoustic features. For the reason of comparison also the SER was also implemented using the two group of features using correlation and PCA. The results are as shown in Fig 7 and 8. From these figures it is clear that the proposed method gives the best MAE for both databases.

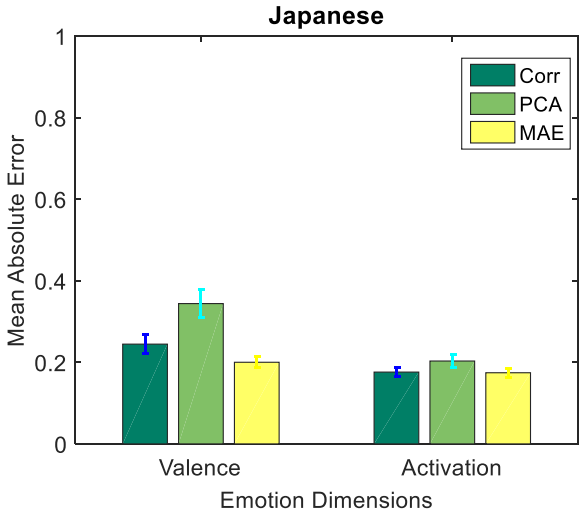


Fig. 7. Comparison between the MAE for Japanese database.

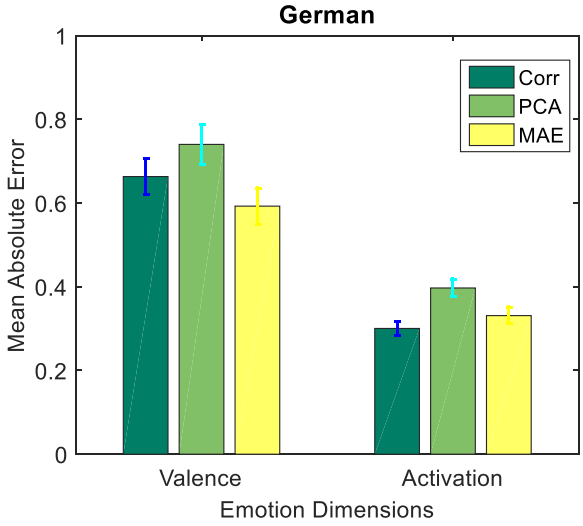


Fig. 8. Comparison between the MAE for German database.

Finally mapping the estimated emotion dimensions values using the third FIS which classifying the utterance into emotion classes. The results for the three methods are listed in Table 4. From this table it is evident that the MAE is the best performance.

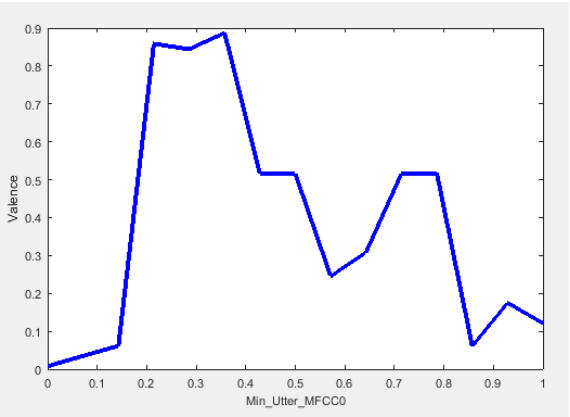
Table 4: The recognition accuracy using different methods.

	CORR	PCA	MAE
Japanese	97%	95%	99%
German	81%	72%	82%

7. DISCUSSION

From the above results it is clear that the proposed method have the ability to find the most related features. Even for the most difficult emotion dimension valence. As listed in table 1, there are the best 10 acoustic features which used to improve the results of recognition accuracy. The results for valence is improved due to that the relation between valence

and acoustic features is non-linear as shown from Fig 9. This figure present the relation between the valence dimension and the best acoustic feature Min_Utter_MFCC_0.



It is clear that this relation is non-linear. Applying the new results for features level and the method for extraction will allow to constructed affective S2ST system. This recognition system is useful for predicting the emotional state of the source language, it can be embedded in many electronic devices includes smartphone and robot devices which can be used for on-line translation between foreigners to overcome the language barrier.

8. CONCLUSTION AND FEUTURE WORKS

This paper propose a feature section method based on FIS to select the most related features for emotion recognition. The proposed method can select not only the features which have a linear but also features with non-liner relation with emotional state. The selected features were used as an input to the proposed speech emotion recognition system. The proposed system consists of two stage the input of the first stage is the selected features and the output is estimated emotion dimension. In the second stage the estimated emotion dimensions were used to recognize the emotional state. Comparing the proposed method with the traditional correlation and PCA. The experimental results shows that the proposed method outperforms the traditional methods for emotion classification.

Our future works is to implement the proposed method for a huge number of acoustic features not only using the MFCC coefficients. Moreover using several statics to extend the domain of used features. Finally using VAD to determine the vowel segment and then apply the feature extraction automatically without any prior knowledge such as the phoneme segmentation information. These results could help to construct online speech emotion recognition system. Which can embedded on the affective speech to speech translation system.

9. REFERENCES

[1] Nakamura, S., "Overcoming the language barrier with speech translation technology," *NISTEP Quarterly Review*, 31, 35-48, 2009.

[2] Shimizu, T., Ashikari, Y., Sumita, E., Zhang, J.S., and Nakamura, S., "NICT/ATR Chinese-Japanese-English Speech-to-Speech Translation System," *Tsinghua Science and Technology, 13, 4, 540-544, 2008.*

[3] Akagi, M. Xiao, H. Elbarougy, R. Hamada, Y. Li, J. "Toward affective speech-to-speech translation: Strategy for emotional speech recognition and synthesis in multiple languages" Proceedings of International Conference (APSIPA2014 ASC), Chiang Mai, December 2014.

[4] R. Jenke, A. Peer, and M. Buss, ''Feature extraction and selection for emotion recognition from EEG,'' IEEE Trans. Affect. Comput. , vol. 5, no. 3, pp. 327–339, Jul./Sep. 2014.

[5] B. Schuller, A. Batliner, S. Steidl, D. Seppi "Recognising realistic emotions and affect in speech: state of the art and lessons learnt from the first challenge" Speech Communication, 53 (9/10) (2011), pp. 1062-1087.

[6] 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.

[7] D. Bitouk, V. Ragini, and N. Ani, "Class-level spectral features for emotion recognition," Journal of Speech Communication, vol. 52, no. 7-8, pp. 613–625, 2010.

[8] M. Grimm, and K. Kroschel, and E. Mower, and S. Narayanan, " Primitives-based evaluation and estimation of emotions in speech," Speech Communication, 49, 787-800 (2007).

[9] H. Gunes, B. Schuller, M. Pantic, and R. Cowie, "Emotion representation, analysis and synthesis in continuous space: A survey," in Proc. IEEE Int. Conf. Autom. Face Gesture Recog., 2011, pp. 827–834.

[10] M. Schroder, and R. Cowie, and E. D.-Cowie, M. Westerdijk, and S. Gielen, "Acoustic Correlates of Emotion Dimensions in View of Speech Synthesis," Proc. Eurospeech 2001 , pp. 87-90 (2001).

[11] H.P. Espinosa, C.A.Reyes-Garca, L.V.Pineda, "Acoustic feature selection and classification of emotions in speech using a 3D continuous emotion model," Biomedical Signal Processing and Control , 7(1), 79-87 (2012).

[12] S. Chen, Q. Jin, X. Li, G. Yang, and J. Xu, "Speech emotion classification using acoustic features," in Chinese Spoken Language Processing (ISCSLP), 2014 9th International Symposium on. IEEE, 2014, pp. 579–583.

[13] R. Elbarougy and M. Akagi. Cross-lingual speech emotion recognition system based on a three-layer model for human perception. 2013 Asia Pacific Signal and Information Processing Association Annual Summit and Conference, pages 1–10, 2013.

[14] R. Elbarougy and M. Akagi, "Speech Emotion Recognition System Based on a Dimensional Approach Using a Three-Layered Model," Proc. Int. Conf. APSIPA ASC, 2012.

[15] Kecman, V., "Learning and Soft Computing: Support Vector Machines, Neural Networks, and Fuzzy Logic Models," MIT Press, (2001)

[16] Wolkenhauer, O., "Data Engineering: Fuzzy Mathematics in Systems Theory and Data Analysis," Wiley, (2001)