Indonesian Journal of Electrical Engineering and Computer Science
Vol. 19, No. 1, July 2020, pp. 196~206
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v19.i1.pp196-206

Fuzzy-based voiced-unvoiced segmentation for emotion recognition using spectral feature fusions

Yusnita Mohd Ali[1], Alhan Farhanah Abd Rahim[2], Emilia Noorsal1[3], Zuhaila Mat Yassin[4], Nor Fadzilah Mokhtar[5], Mohamad Helmy Ramlan[6]
[1,2,3,4,5]Faculty of Electrical Engineering, Universiti Teknologi MARA, Malaysia
[6]ASE Electronics (M) Sdn. Bhd. Phase 4, Malaysia

Article Info

Article history:

Received Okt 24, 2019
Revised Dec 10, 2019
Accepted Feb 2, 2020

Keywords:

Emotion recognition
Fuzzy logic
Linear prediction coefficients
Mel-frequency cepstral coefficients
Short-time energy
Zero-crossing rate

ABSTRACT

Despite abundant growth in automatic emotion recognition system (ERS) studies using various techniques in feature extractions and classifiers, scarce sources found to improve the system via pre-processing techniques. This paper proposed a smart pre-processing stage using fuzzy logic inference system (FIS) based on Mamdani engine and simple time-based features i.e. zero-crossing rate (ZCR) and short-time energy (STE) to initially identify a frame as voiced (V) or unvoiced (UV). Mel-frequency cepstral coefficients (MFCC) and linear prediction coefficients (LPC) were tested with K-nearest neighbours (KNN) classifiers to evaluate the proposed FIS V-UV segmentation. We also introduced two feature fusions of MFCC and LPC with formants to obtain better performance. Experimental results of the proposed system surpassed the conventional ERS which yielded a rise in accuracy rate from 3.7% to 9.0%. The fusion of LPC and formants named as SFF LPC-fmnt indicated a promising result between 1.3% and 5.1% higher accuracy rate than its baseline features in classifying between neutral, angry, happy and sad emotions. The best accuracy rates yielded for male and female speakers were 79.1% and 79.9% respectively using SFF MFCC-fmnt fusion technique.

Corresponding Author:

Yusnita Mohd Ali,
Faculty of Electrical Engineering,
Universiti Teknologi MARA, Cawangan Pulau Pinang
13500 Permatang Pauh, Pulau Pinang, Malaysia.
Email: yusnita082@uitm.edu.my

1. INTRODUCTION

Emotions are complex manifestations of psychological and physiological phenomena of human beings that involve psychological arousal, expressive behaviours and conscious experience [1]. It can be conveyed through human actions, facial expressions and voices. The expression of emotion via speech signal can be regarded as the most natural, fast and efficient communication means to tell the other party of what is inside of one's heart. There has been a large body of research in emotion recognition using physiological signals [2, 3], facial images and videos [4, 5] including human speech [6-9] to correlate with emotions for various applications such as in security system, classroom pedagogy, customer service call centres and job matching marketplace through phone interviews.

Emotional states can be described using three- or simply two-dimensional model namely valence, arousal and control. However the third dimension is rarely used since it occupies only a narrow range in the model. Valence relates to a continuum range of positive emotion (pleasant) to negative emotion (unpleasant), while arousal or activation is characterized by the intensity of emotional state from energized, excited and alert to calm, drowsy and peaceful [10]. However, El Ayadi, Kamel and Karray [11] in their survey

did not find an agreement among researchers to correlate acoustic speech features to these dimensions. Even so, there is certain correlations between the physiological process in the sympathetic and parasympathetic autonomic nervous systems of emotion production mechanism to speech production system which affect pitch, timing, voice quality and articulation of speech signal. In pattern recognition, categorical approach of emotional states are more useful to discriminate between emotional classes. A large number of emotional categories were proposed by Whissel, Plutchik and Cowie in the 20th and 21st centuries. Nevertheless, the current emotion recognition system (ERS) can only detect between two to eight classes of emotional states with reasonably good results.

Speech is complex signal and non-stationary in nature. As such, speech is pre-processed in frame-wise of 20 ms to 50 ms length to ensure stationary property [12] and the choice of suitable frame or window size is also vital to ensure appropriate spectral estimate of each frame [13]. There were also disputes of whether local speech features, calculated from each frame, or global features taking statistics of all local speech features of an utterance was better indicator for emotion cues. There had been trade-off between these two techniques wherein the latter performed better in terms of accuracy rate and processing time as compared to the local features [11]. However, it could not differentiate emotions of similar arousal such as anger and joy and losing of temporal information. In this work, we preferred to take a fixed-length utterance-level statistics, namely mean values while ignoring the temporal dynamics attributes such as in [14]. To date, various methods have been proposed that focus on feature extraction and classification stages. There is no consensus of which features or classifiers are the best methods for emotion recognition task because recognizing human emotion is a difficult problem having high nonlinearity and variability. Hence, it is still an open question to explore the features and classifiers that can accomplish the best results.

Spectral-based features extracted from short-time frame signal were more popularly used, for instance linear prediction coefficients (LPC), log-frequency power coefficients (LFPC) and cepstral-based features such as linear prediction cepstral coefficients (LPCC) and Mel-frequency cepstral coefficients (MFCC). Bandela and Kumar proposed a fusion of Teager energy operator (TEO) with MFCC [15] and a fusion of TEO with LPC [16] on Berlin Emotional Acted Database (EMO-DB) for five emotional states classification using gaussian mixture model (GMM). They found accuracy rates of 93.3% and 88.0% using the proposed features, surpassing the baseline features of MFCC and LPC by 6.7% and 9.3% respectively. Another recent work using Amritaemo database [17] compared the performance of artificial neural networks (ANN) and K-nearest neighbours (KNN) to classify happy, angry and sad in Telegu and Tamil languages and obtained 76% and 75% accuracy rates using a combination of Hurst parameters and linear predictive cepstral coefficients (LPCC).

In a recent article, Jiang et al. [18] investigated the effectiveness of different classifiers such as KNN, GMM and support vector machine (SVM) for detecting depression from positive, negative and neutral speech emotions from 170 subjects. A highly dimensional features using prosodic, voice quality and spectral features were used. It was found that SVM surpassed the other two with 65.7% accuracy rate. Among three types of speech, picture description yielded the best input for male while interview speech was the one for female speakers. In most past studies, researchers used a combination of the aforementioned features to obtain the best accuracy of emotion detection [7, 11, 19]. Although various techniques have been investigated in feature extraction and classification stage in speech processing [20], there is scarce resources on pre-processing stage in ERS. Since frame blocking is necessary in speech processing, selecting appropriate frames before feature extraction is important task to improve the performance of ERS because not all frames contain emotion attribute especially when global features of an utterance is taken such in this paper.

One of the simple selection criteria is to classify the speech frames as voiced (V) or unvoiced (UV). The performance of ERS on compressed speech was investigated in [9] using seven digital telecommunication codecs based on adaptive differential pulse code modulation (ADPCM) and Analysis-by-Synthesis. V and UV segments of speech data was classified using autocorrelation method in Praat software tool. Several parameters were extracted using MFCC, Bark spectral energy, noise measures, and nonlinear dynamics features from V and UV speech frames to build two recognition models separately based on GMM. This research concluded that the UV-based ERS produced a significant degradation in accuracy while just a little drop in accuracy was noticed for V-based ERS due to compression by the codecs on EMO-DB and enterface05 emotional speech databases.

Extracting paralinguistics traits that are related to emotion recognition using prosodic features such as formants and fundamental frequency (F0) would be incorrect for unvoiced speech because these features were valid for voiced speech [21]. In [22], V and UV segments were determined using F0 estimation and different sets of features were extracted from these types of segments to recognize emotions in continuous scale instead of discrete emotions using three dimensional values of valance, activation and dominance. The authors also suggested optimizing features using particle swarm optimization (PSO) to reduce the feature space dimensionality from 830 of V-segments and 710 of UV-segments to only 40 optimal features selected by PSO separately for V and UV segments. Overall, this had resulted better classification rates of emotional

dimensions.Based on the limitation of previous studies, this paper proposes a development of ERS which employs a smart pre-processing stage to identify a frame as V-frame or UV-frame using fuzzy logic inference system (FIS) modelled with simple time-based features i.e. zero-crossing rate (ZCR) and short-time energy (STE). Since voiced speech has stronger emotional trait than unvoiced, the proposed FIS V-UV segmentation will output only V-frames to feature extraction stage for more efficient and accurate performance. This study uses an acted speech corpora consisted of four basic emotional states namely neutral, angry, happy and sad. Apart from baseline features such as MFCC and LPC, this paper proposes feature fusions of MFCC and LPC with the first five formants.

2. RESEARCH METHOD

The development of fuzzy inference system (FIS) for voiced-unvoiced segmentation in ERS follows the steps described in Figure 1. Basically, the system consists of pre-processing and feature extraction in the front-end while in the back-end, it involves signal modelling and classification stage of the unknown speech samples. A modification is made by classifying the pre-processed frames into either voiced (V) or unvoiced (UV) frame before extracting the acoustic features. The development stages are explained in this section. The hypotheses presumes that by incorporating FIS for identifying V-frame or UV-frame the system efficacy could be improved in terms of accuracy and processing time wherein only voiced frames are taken for further processing after FIS decision has been made.

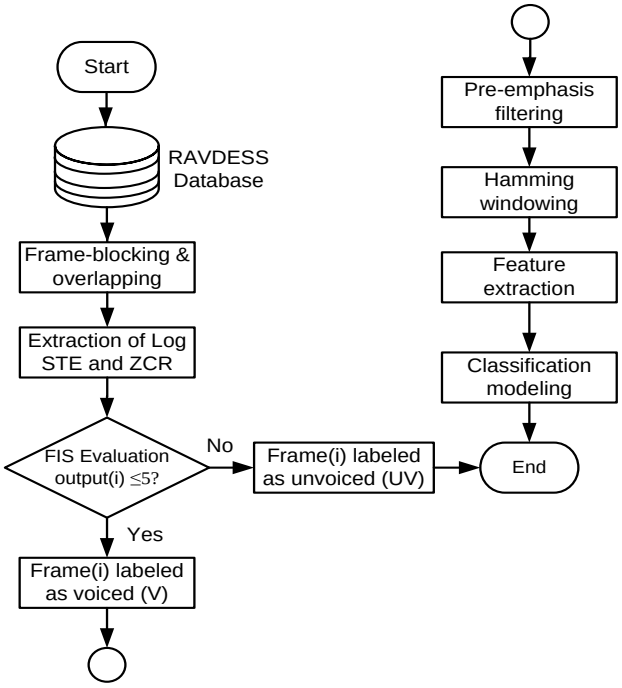


Figure 1. Emotion recognition with fuzzy inference system for voiced-unvoiced segmentation

2.1. Emotional database

The speech database used in this paper was collected from The Ryerson Audio-Visual Database of Emotion Speech and Song (RAVDESS). This database [23] elicited speech from 24 professional actors consisted of an equal number of male and female speakers respectively. The actors vocalized two lexically-matched statements in North American accent. The database was available in three modality i.e. audio only, audio-video and video only. Only audio format .wav files were taken which were sampled at 48 kHz and quantized at 16 bit. The sampling frequency was downsampled to 16 kHz. The speakers uttered two sentences namely "Kids are talking by the door" and "Dogs are sitting by the door" in four acted emotionally conditions such as angry, sad, happy and neutral emotions. There existed two different emotional intensity while speaking the sentences labelled as normal intensity and strong intensity. However, neutral emotion was only available with normal intensity. As such, we only took strong intensity for this study. Each statement was repeated two times for each emotion. Table 1 summarizes the number of samples extracted for this research use.

Table 1. RAVDESS emotional database samples

Emotional State	Gender		No. of utterances
	Male	Female	
Angry	48	48	96
Sad	48	48	96
Happy	48	48	96
Neutral	48	48	96
Total	192	192	384

2.2. Fuzzy inference system segmentation

The fuzzy inference system (FIS) is designed using MATLAB fuzzy logic toolbox GUIs as depicted in Figure 2 having two inputs and one output. The most widely accepted fuzzy inference method for capturing expert knowledge is Mamdani coined by Ebrahim Mamdani of London University in 1975 to control a steam engine and boiler combination. Despite the fact that Mamdani-type fuzzy inference entails a substantial computational burden, it allows the description of expertise in more intuitive and human-like manner [24, 25]. There are many ways to assign membership values and functions to fuzzy variables and the one that is proposed in this paper is by inductive reasoning using values derived from experimental data.

The designed FIS V-UV using membership functions of short-time energy (STE) and zero-crossing rate (ZCR) as inputs are shown in Figure 2. Triangular and trapezoidal membership functions are adopted as these functions are practical and one of the simplest linear-fit functions. The local statistical thresholds for STE and ZCR experimental data as proposed in this paper are mathematically given as in (1) and (2) respectively.

$$\theta(k,s) = \begin{bmatrix} \theta_{11} & \theta_{12} & \theta_{13} \\ \theta_{21} & \theta_{22} & \theta_{23} \\ \theta_{31} & \theta_{32} & \theta_{33} \\ \vdots & \vdots & \vdots \\ \theta_{k1} & \theta_{k2} & \theta_{k3} \end{bmatrix} \quad (1)$$

$$(k,s) = \begin{bmatrix} \varphi_{11} & \varphi_{12} & \varphi_{13} \\ \varphi_{21} & \varphi_{22} & \varphi_{23} \\ \varphi_{31} & \varphi_{32} & \varphi_{33} \\ \vdots & \vdots & \vdots \\ \varphi_{k1} & \varphi_{k2} & \varphi_{k3} \end{bmatrix} \quad (2)$$

where $\theta(k,s)$ and $\varphi(k,s)$ are the local thresholds of STE and ZCR. The first subscript, k denotes the speech number that is contained in the current directory, and the second subscript s represents the statistical descriptors s which takes value 1 for minimum (min), 2 for median (med) and 3 for maximum (max). Finally, global statistical thresholds (GSTs) can be calculated from the matrices of the local thresholds as given in (3) and (4).

$$\theta_G(s) = \begin{bmatrix} \min\theta_{min} & \text{med}\,\theta_{min} & \max\theta_{min} \\ \min\theta_{med} & \text{med}\,\theta_{med} & \max\theta_{med} \\ \min\theta_{max} & \text{med}\,\theta_{max} & \max\theta_{max} \end{bmatrix} \quad (3)$$

$$\varphi_G(s) = \begin{bmatrix} \min\varphi_{min} & \text{med}\,\varphi_{min} & \max\varphi_{min} \\ \min\varphi_{med} & \text{med}\,\varphi_{med} & \max\varphi_{med} \\ \min\varphi_{max} & \text{med}\,\varphi_{max} & \max\varphi_{max} \end{bmatrix} \quad (4)$$

where θ_G (k,s) and are φ_G (k,s) the global thresholds of STE and ZCR and s is the statistical descriptors as described above.

Table 2 shows the resulted global thresholds calculated from all speech samples of both gender. Since the histograms shown for STE and ZCR have skewed distributions, median values are calculated instead of mean. The working of the V-UV segmentation using FIS can be demonstrated by feeding some input combination to predict the output of a frame either V or UV. Figure 3(a) is an example of unvoiced detected frame, while Figure 3(b) is an example of voiced detected frame. The fuzzy rules which are established in this design of FIS V-UV are tabulated in Table 3 after some justification of expert knowledge of human speech production. Four rules consist of two antecedents and one consequent were formulated in conditional statements if-then rules based on the studied characteristics of V-UV on STE and ZCR using rigorous analysis and observation. The first and second rules are highly correlated with the past studies [26, 27]. The third and fourth rules are formulated to save most part of the speech since from the proposed GSTs analysis that can be referred from Figure 2(b) and Figure 2(c). The range of the large ZCR subset extends the largest.

On the other hand, the small and median subsets of STE occupy more spaces than the large subset of STE. Hence, it is proposed that median STE with small ZCR would make a frame label as V and large STE with median ZCR would also make a frame as V. By these rules, it means roughly 75% of the speech has possibility to be V frames.

To demonstrate the resulted voiced and unvoiced frames using the above formulated rules, Figure 4 shows a mixture of characteristics of STE and ZCR plotted against frame number of a speech signal of a female speaker having sad emotion identified as 03-01-04-02-01-01-16. This speech consisted of 117 frames, 42 of which were labeled as UV and the remaining 75 were labeled as V after FIS evaluation. The voiced frames (labeled as V-frames) would to be preserved for the next step of speech processing. In addition to this, Figure 5 displays the short-time speech waveforms of a voiced frame, the 12th frame, and an unvoiced frame, the 6th frame from the previous result. The values of STE and ZCR for the 12th frame of this signal are found to be 17.70 and 32 respectively, as the V-UV segmentation FIS decided it to be a voiced-type after FIS evaluation. Whilst the values of STE and ZCR for the 6th frame of this signal are found to be -2.56 and 226 respectively, as the V-UV segmentation FIS decided it to be an unvoiced-type after FIS evaluation.

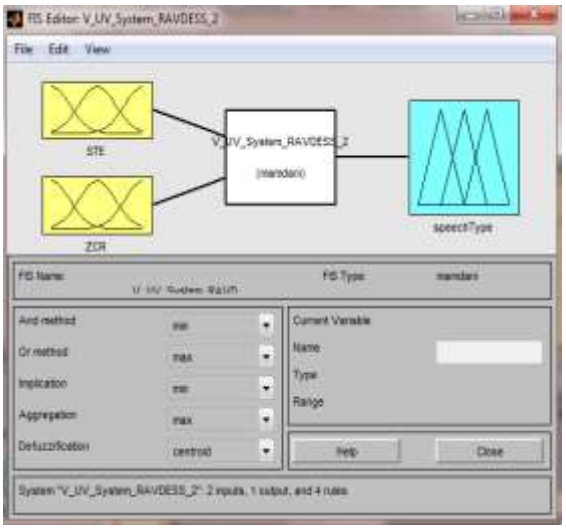


Figure 2(a). FIS segmentation system

Figure 2(b). Membership functions to fuzzify

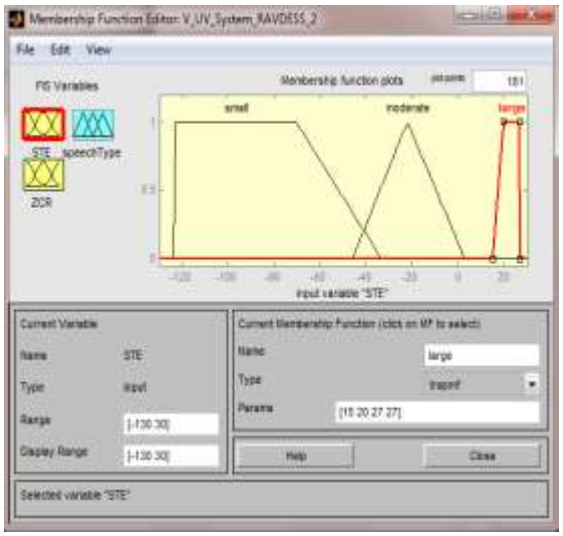


Figure 2(c). Zero-crossing rate inputs

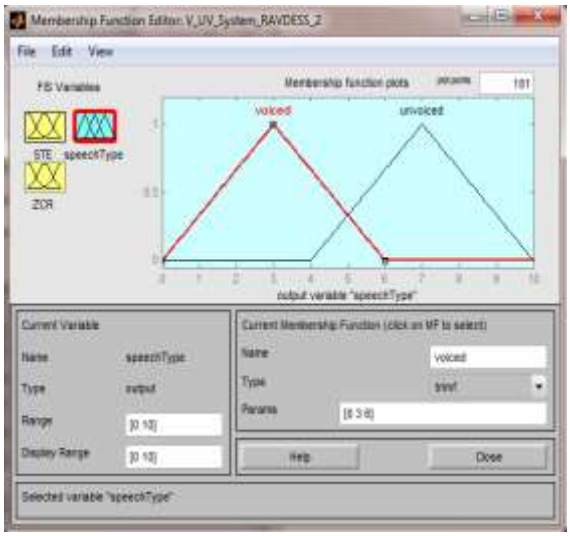


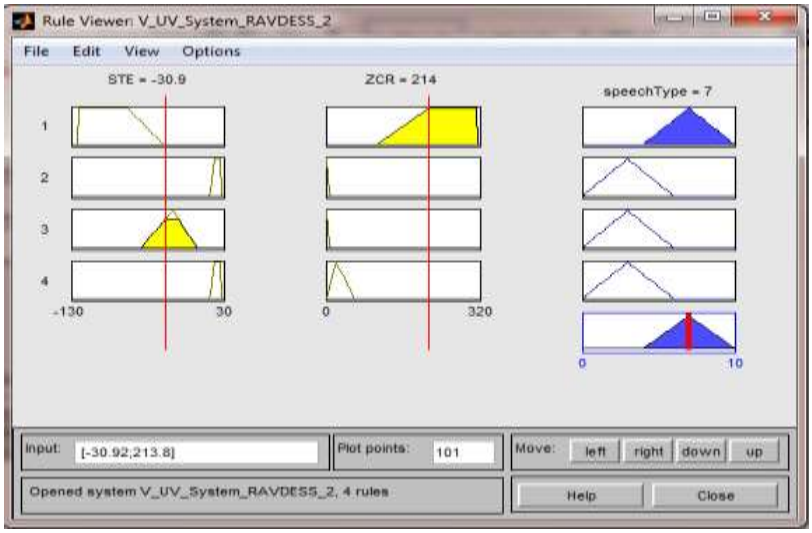
Figure 2(d). Membership functions to de-fuzzify the type of speech frames

Table 2. Global statistical thresholds (GSTs) for STE and ZCR

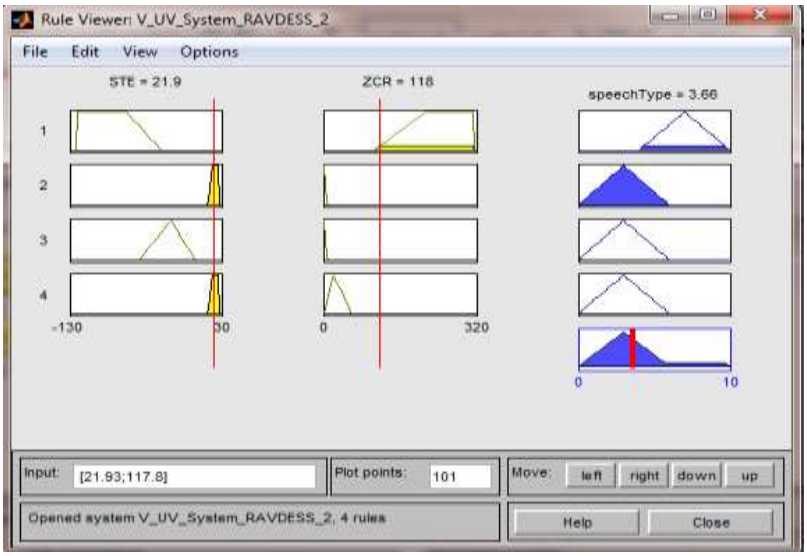
STE global threshold (θG)	Stat	STE local threshold (θ)	ZCR global threshold (φG)	Stat	ZCR local threshold (φ)
	min	-123		min	0
Min	med	-70	Min	med	0
	max	-33		max	7
	min	-46		min	9
Med	med	-22	Med	med	35
	max	3		max	72
	min	14.8		min	103
Max	med	19.5	Max	med	213
	max	26.8		max	316

Table 3. Fuzzy rules to build voiced-unvoiced Fuzzy inference system

Rule No.		STE	Operator	ZCR		Output
1		Small	or	large		UV
2	IF	Large	or	small	THEN	V
3		median	and	small		V
4		Large	and	median		V



(a)



(b)

Figure 3. Rule viewer editor to show fuzzy rules evaluation on a given input combination for (a) UV-frame, (b) V-frame

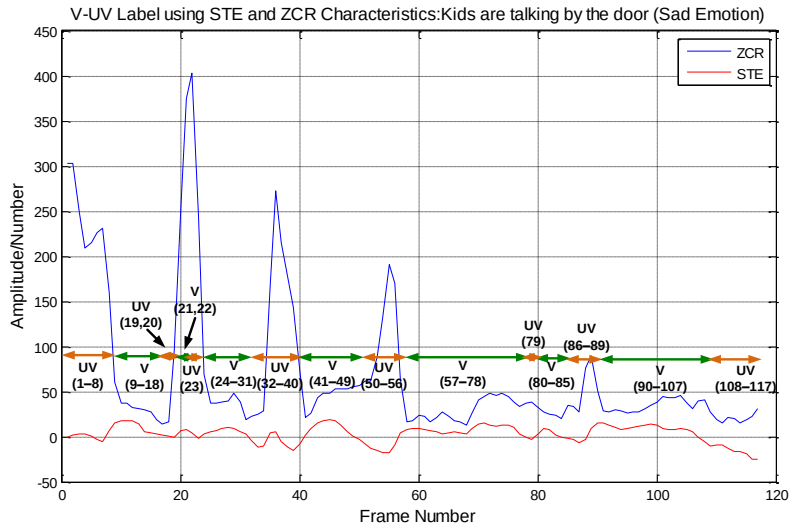


Figure 4. Classification of frames into V-UV class based on the profiles of STE and ZCR

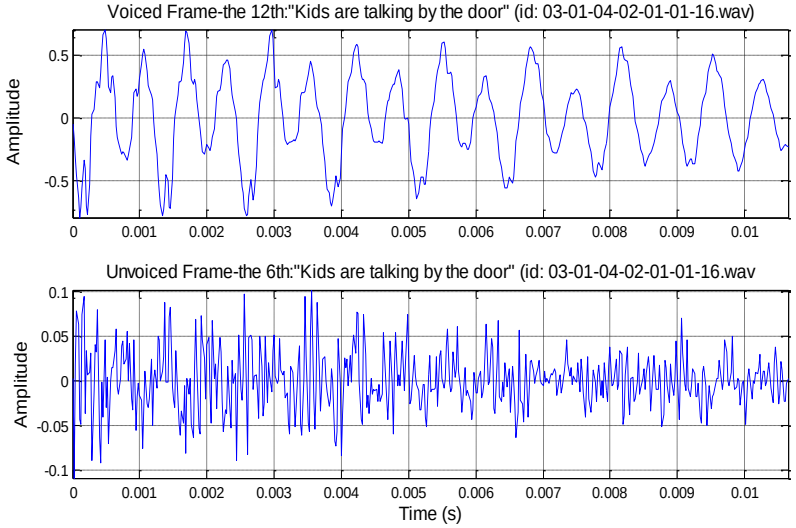


Figure 5. Voiced frame (upper figure) and an unvoiced frame (lower figure) after FIS V-UV segmentation

3. RESULTS AND ANALYSIS

In this paper, we employed a simple classification technique i.e. KNN [28] to classify an unknown speech into four emotional classes namely, Class 1: Neutral, Class 2: Angry, Class 3: Happy and Class 4: Sad using different distance measures and number of nearest neighbours. The rule used here to classify the sample was nearest which means majority rule with nearest point tie-break to make the class prediction.

The popularly used parametric features such as LPC and MFCC were adopted as the inputs to KNN models. Another important features used was formant frequency or simply formants which can be extracted from LPC spectrum as peaks starting from the highest energy represented as the first formant F1, followed by the second formant F2, and so forth up to the fifth formant F5 in order. The mixture of spectral-based features (LPC and MFCC) and prosodic features (formants) were tested to compare the performance of the proposed FIS V-UV segmentation into this ERS system. The system were analysed as follows. Firstly, the performance of the proposed V-UV segmentation is shown by comparing the overall accuracy rates of ERS system with and without employing FIS V-UV segmentation algorithm. In addition, the formants were fused with the spectral features to generate two new feature vectors namely spectral feature fusion of MFCC and formants (SFF MFCC-fmnt) and LPC mixed with formants (SFF LPC-fmnt). The combinations of formants with MFCC and LPC were motivated by previously obtained poor results of using formants alone. The chosen parameters

of KNN classifier were K = 2 and cityblock distance metric for all acoustic features except for MFCC-based features for female speakers which adopted correlation metric. Table 4 shows the reduction in number of frames to be processed for feature extraction using FIS implementation. The proposed segmentation scheme was able to save 31.7% and 36.4% of frames for gender sensitive ERS prior to feature extraction and emotion classification stages.

Figure 6 shows the results using baseline MFCC and LPC as well as their fusion features with formants for both male and female speakers. It was found that the mixture of MFCC and formants did not give much rise in accuracy of emotion detection both in male and female speech. The change was less than 1% after FIS V-UV segmentation. However, there was a noticeable increase in accuracy with and without FIS V-UV segmentation for SFF LPC-formants from its baseline features by approximately 1.3% to 5.1%.

Table 4. Reduction rate of number of frames using FIS V-UV segmentation

Frame type	No of frames	
	Male	Female
Voiced (V)	91,021	87,055
Unvoiced (UV)	42,269	49,837
Reduction rate (%)	31.7	36.4

Table 5. Accuracy rate changes using FIS V-UV ERS

Speech features	Classification rate (%)	
	Male	Female
MFCC	-2.0	+1.1
LPC	+6.4	+9.0
SFF MFCC-fmnt	-0.5	+2.9
SFF LPC-fmnt	+3.7	+7.5

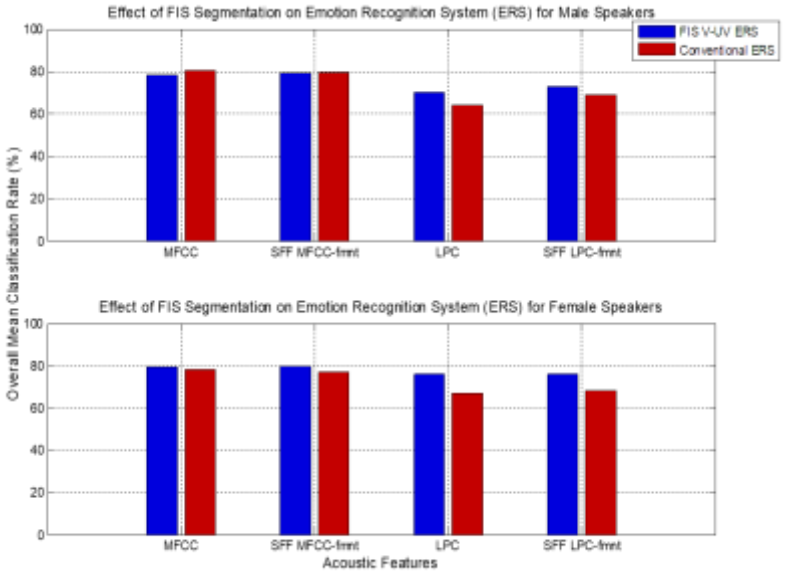


Figure 6. Comparison of performance between conventional ERS and ERS employing FIS V-UV segmentation for different acoustic features

Comparing the effect of FIS V-UV segmentation on ERS systems, we succeeded to show that the proposed segmentation using fuzzy logic worked effectively on most of the feature vectors based on KNN classifiers. Table 5 tabulates the accuracy rate changes from conventional ERS to FIS V-UV ERS using all four feature vectors extracted from male and female emotional speech. The positive sign indicates an increase while a negative sign indicates a decrease in overall-class mean classification rate. The success rates were consistently promising using LPC-based features namely increment between 3.7% and 9.0% for male and female speech for detecting four-class emotional state (overall class performance). Female have greater rise in classification rates than male for all types of feature vectors. Overall, the highest accuracy rates for male and female gender-sensitive ERS were 79.1% and 79.9% respectively as a result of FIS V-UV implementation.

Finally, we show the performance (mean classification rate) of individual class emotional state from voiced segments of male and female speakers as depicted in Figure 7. Again, most features yielded the best results for emotional-free speech, namely neutral i.e. 91.1% for male speech (MFCC) and 91.6% for female speech (LPC and SFF MFCC-fmnt). This was followed by angry, happy and sad in the same sequence as obtained previously. Sad emotion being the lowest class correctly recognized, appeared stronger in female speech 73.7% as compared to 64.7% in male speech using MFCC.

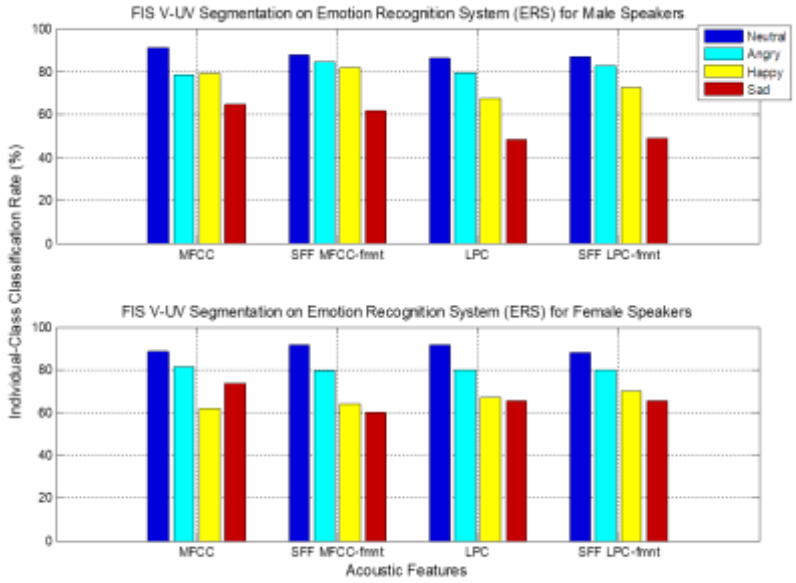


Figure 7. Individual-class performance of FIS V-UV ERS using different acoustic features

4. CONCLUSION

In this paper, we introduced an efficient technique to improve emotion recognition system (ERS) using fuzzy expert system for segmenting speech into either voiced or unvoiced (V-UV) class. The system was developed using global statistical thresholds (GSTs) of time-based features namely, log of short-time energy (STE) and zero-crossing rate (ZCR) and Mamdani inference method as the engine. Speech utterances from male and female speakers obtained from RAVDESS database were extracted using the spectral features, MFCC and LPC and fused with the prosodic features, five formant frequencies in an attempt to improve the recognition rate. We succeeded to show that the proposed system, FIS V-UV segmentation when employed in ERS resulted a better performance over the conventional ERS which yielded increment of accuracy rate from 3.7% to 9.0% on LPC-based features using KNN classification algorithm. Comparing between genders, it shows that FIS V-UV segmentation has better classification rates on female database for all four tested feature vectors i.e. MFCC, LPC, SFF MFCC-fmnt and SFF LPC-fmnt as compared to male database.

The fusion of LPC and formants named as SFF LPC-fmnt also showed a promising results of 1.3% to 5.1% higher in overall classification rate than its baseline features for this four-class problem. It is also found that emotional-free speech, neutral was classified most correctly than other emotional speech such as angry, happy and sad with sad classified the lowest rate. This can indirectly mean that emotion traits in speech could degrade the automatic speech recognition systems in general. As final conclusion, the proposed FIS V-UV segmentation achieved the highest overall class mean classification rates of 79.1% and 79.9% for male- and female-sensitive ERS respectively using SFF MFCC-fmnt feature fusion technique.

REFERENCES

[1] D. G. Myers, "Theories of emotion," vol. 500, 2004.
[2] J. R. Zhuang, *et al.*, "Two-dimensional emotion evaluation with multiple physiological signals," *Advances in Affective and Pleasurable Design. AHFE 2018. Advances in Intelligent Systems and Computing*, vol. 774, 2019.
[3] C. Li, *et al.*, "Emotion recognition of human physiological signals based on recursive quantitative analysis," *Tenth International Conference on Advanced Computational Intelligence*, pp. 217-223, 2018.

[4] B. M. Ghandi, R. Nagarajan, and H. Desa, "Real-time system for facial emotion detection using GPSO algorithm," *IEEE Symposium on Industrial Electronics and Applications*, pp. 40-45, 2010.
[5] F. Z. Salmam, A. Madani, and M. Kissi, "Emotion recognition from facial expression based on fiducial points detection and using neural network," *Int. J. of Elect. and Comp. Enginee.*, vol. 8, p. 52, 2018.
[6] P. Heracleous, *et al.*, "Speech emotion recognition in noisy and reverberant environments," *Seventh International Conference on Affective Computing and Intelligent Interaction*, pp. 262-266, 2018.
[7] L. Kerkeni, *et al.*, "A review on speech emotion recognition: Case of pedagogical interaction in classroom," *International Conference on Advanced Technologies for Signal and Image Processing*, pp. 1-7, 2017.
[8] T. Saste and S. M. Jagdale, "Emotion recognition from speech using MFCC and DWT for security system," *International conference of Electronics, Communication and Aerospace Technology*, pp. 701-704, 2017.
[9] N. Garcia, *et al.*, "Automatic emotion recognition in compressed speech using acoustic and non-linear features," *20th Symposium on Signal Processing, Images and Computer Vision*, pp. 1-7, 2015.
[10] L. He, "Stress and emotion recognition in natural speech in the work and family environments," *School of Electrical and Computer Engineering Science, Engineering and Technology Portfolio*, p. 218, 2010.
[11] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572-587, 2011.
[12] S. Furui, "Fifty years of progress in speech and speaker recognition," *ECTI Transaction on Computer and Information Technology*, vol. 1, pp. 64-74, 2005.
[13] M. Ghai, S. Lal, S. Duggal, and S. Manik, "Emotion recognition on speech signals using machine learning," *International Conference on Big Data Analytics and Computational Intelligence*, pp. 34-39, 2017.
[14] Z. Wang and I. Tashev, "Learning utterance-level representations for speech emotion and age/gender recognition using deep neural networks," *IEEE Int. Conference on Acoustics, Speech and Signal Processing*, pp. 5150-5154, 2017.
[15] S. R. Bandela and T. K. Kumar, "Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC," *International Conference on Computing, Communication and Networking Technologies*, pp. 1-5, 2017.
[16] S. R. Bandela and T. K. Kumar, "Emotion Recognition of Stressed Speech Using Teager Energy and Linear Prediction Features," *IEEE 18th International Conference on Advanced Learning Technologies*, pp. 422-425, 2018.
[17] S. Renjith and K. G. Manju, "Speech based emotion recognition in Tamil and Telugu using LPCC and hurst parameters: A comparitive study using KNN and ANN classifiers," *International Conference on Circuit, Power and Computing Technologies*, pp. 1-6, 2017.
[18] H. Jiang, *et al.*, "Investigation of different speech types and emotions for detecting depression using different classifiers," *Speech Communication*, vol. 90, pp. 39-46, 2017.
[19] A. Khalil, *et al.*, "Anger Detection in Arabic Speech Dialogs," *International Conference on Computing Sciences and Engineering*, pp. 1-6, 2018.
[20] T. S. Gunawan, N. A. M. Saleh, and M. Kartiwi, "Development of quranic reciter identification system using MFCC and GMM classifier," *International Journal of Electrical and Computer Engineering*, vol. 8, pp. 372-378, 2018.
[21] Y. Li, et al., "Classification of voices that elicit soothing effect by applying a voiced vs. unvoiced feature engineering strategy," *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 2489-2493, 2016.
[22] M. Hric, M. Chmulík, I. Guoth, and R. Jarina, "SVM based speaker emotion recognition in continuous scale," *25th International Conference Radioelektronika*, pp. 339-342, 2015.
[23] S. R. Livingstone, K. Peck, and F. A. Russo, "RAVDESS: the ryerson audio-visual database of emotional speech and song," *Annual Meeting of the Canadian Society for Brain, Behaviour and Cognitive Scienc*e, Kingston, 2012.
[24] L. A. Zadeh, "Outline of a New Approach to the Analysis of Complex Systems and Decision Processes," *Systems, Man and Cybernetics, IEEE Transactions on*, vol. SMC-3, pp. 28-44, 1973.
[25] A. Amindoust, et al., "Sustainable supplier selection: A ranking model based on fuzzy inference system," *Applied Soft Computing*, vol. 12, pp. 1668-1677, 2012.
[26] B. S. Atal and L. R. Rabiner, "Pattern recognition approach to voiced-unvoiced-silence classification with applications to speech recognition," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, pp. 201-212, 1976.
[27] M. Malcangi, "Softcomputing approach to segmentation of speech in phonetic units," *International Journal of Computer and Communications*, vol. 3, pp. 41-48, 2009.
[28] M. A. Yusnita, *et al.*, "Malaysian English accents identification using LPC and formant analysis," *IEEE International Conference on Control System, Computing and Engineering*, pp. 472-476, 2011.

BIOGRAPHIES OF AUTHORS

Yusnita Mohd Ali is a senior lecturer at the Faculty of Electrical Engineering, Universiti Teknologi MARA, Penang Campus, Malaysia. She received her PhD Degree in Mechatronic Engineering from Universiti Malaysia Perlis in 2014 specializing in Audio/Acoustic Engineering. She was conferred with a Master Degree in Electronics System Design Engineering from University Sains Malaysia in 2004. She completed her Bachelor Degree in Electrical & Electronics Engineering from the same university in 1998. Her field of interest includes speech processing, speech analysis, human-machine interaction, brain-machine communication and artificial intelligence.

Mohamad Helmy Bin Ramlan is currently working at ASE Electronic (M) Sdn. Bhd. as Test Product Engineer. He is responsible for product yield, quality, test program maintenance/release and customer support on electrical test related activities. He received his Diploma in Electrical Engineering (Instrumentation) from UiTM Pasir Gudang, Johor in 2015. In same year, he pursued his study in Bachelor Degree in Electrical & Electronics Engineering in UiTM Pulau Pinang and ted it in 2018.

Zuhaila Mat Yasin graduated from Universiti Sains Malaysia, with honours degree in Electrical & Electronics Engineering in 1998. She obtained her MSc degree in 2008 and PhD degree in 2014 from Universiti Teknologi MARA, Malaysia. She currently lectures at Universiti Teknologi MARA, Malaysia. Her research interest includes power system planning, power system stability, distributed generation, renewable energy and artificial intelligence.

Alhan Farhanah Abd Rahim is a senior lecturer attached to the Faculty of Electrical Engineering, Universiti Teknologi MARA, Penang Campus, Malaysia. She received her PhD degree in Solid State Physics from Universiti Sains Malaysia in 2014 with her Thesis entitle: Studies of Si, Ge and ZnO low dimensional structures synthesized by photo-electrochemical and plasma assisted techniques for sensing applications. Her research interests include semiconductor fabrication, nanostructure from group IV and III-V for optoelectronic application and gas sensor.

Emilia Noorsal is a senior lecturer at the Universiti Teknologi MARA Penang Campus, Malaysia. In April 2014, she obtained her PhD in biomedical engineering from Institute of Microelectronics, Ulm, Germany. Her research interests include digital design circuit in ASIC, FPGA, mixed-signal circuit design, power electronics and electronics for biomedical applications

Nor Fadzilah Mokhtar is a senior lecturer attached to the Faculty of Electrical Engineering, Universiti Teknologi MARA Penang Campus, Malaysia. She received her MSc. degree in Electronics System Design Engineering from Universiti Sains Malaysia in 2004. Her research interests include digital design circuit in ASIC, embedded system, advanced control system, artificial intelligence, and power electronics.