Received: June 15, 2020. Revised: July 2, 2020.

International Journal of
Intelligent Engineering & Systems

http://www.inass.org/

Speech Emotion Recognition Using MELBP Variants of Spectrogram Image

Suhaila N. Mohammed[1,2]* Alia K. Abdul Hassan[1]

[1]*Computer Sciences Department, University of Technology, Baghdad, Iraq*
[2]*Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq*
* Corresponding author's Email: suhailan.mo@sc.uobaghdad.edu.iq

Abstract: Speech emotion recognition finds many applications in the daily life like conversational agents, human robot interaction, call centres etc. However; the task of emotion recognition from speech signal is not trivial due to the difficulty in determining the effective feature set that can recognize the emotion conveyed within the signal in an accurate manner. Image processing techniques are exploited in this paper to solve speech emotion recognition problem. After converting the signal into 2D spectrogram image representation, four forms of Extended Local Binary Pattern (ELBP) are generated to serve as a source for feature extraction stage. The histograms of multiple blocks from ELBP variants are computed and fed to Deep Belief Network (DBN) for classification purpose. Different tests were performed using Surrey Audio-Visual Expressed Emotion (SAVEE) database and the achieved results showed that when using combined vectors of MELBP, the system gives the best accuracy which is 72.14%. The achieved result outperforms state-of-the-art results on the same database.

Keywords: Speech emotion, Spectrogram image, Multi-block extended local binary pattern (MELBP), Deep belief network (DBN), Short term fourier transform (STFT).

1. Introduction

Speech Emotion Recognition (SER) refers to the extraction of feelings from speech signals. Different applications, relying on the user's emotional state, can benefit from SER systems such as human-robot interaction, pain and lying detection, computer-based tutorial systems, and movie or music recommendation systems [1, 2].

In general, a SER system uses a classifier to recognize the emotion from the feature vector that is extracted from the speech signal. A SER system must be robust against speaking rate and speaker style. It means special characteristics like differences in age, gender and culture should not impact on the performance of the SER system. Many researchers are working in this field to give the machine the intelligence in understanding the emotion state using the speech signal of the user. In SER, the investigation and extraction of relevant and discriminatory features is a difficult task [3].

There are different methods used for speech feature extraction such as continuous-based features, spectral-based speech features and digital image processing techniques. Many researchers believe that effective continuous features such as pitch and energy reflect most of an utterance's emotional content, for example, the speaker's arousal state influences the overall energy, and the length and duration of speech pauses. Some of famous continuous acoustic features are energy related features, pitch, formants and timing related features. Pitch is a fundamental property of the speech signal. The pitch describes the highness and lowness of tone in the speech. Pitch features increased with high-arousal emotions such as happy and surprise emotions while decreased with low-arousal emotions such as sad and fear. In phonetics, formant essentially means the acoustic resonance of the vocal tract of the human [4]. They can be extracted by finding the amplitude peaks in the frequency spectrum of the speech. Timing-related features provide information about the distribution of duration-related parameters such speech rate, the

percent between voiced and unvoiced parts. Timing features increase with high-arousal emotions and decrease with low-arousal emotions [5].

On the other hand; it is identified that the emotional state of an utterance has an effect on the distribution of the spectral energy throughout the frequency range of the speech signal and the effective features can be extracted from that domain such as Mel frequency Cepstral Coefficients (MFCCs) [6]. For instance, it is found that speech signal with happiness emotion own high energy at the high frequency range while the signal with the sadness emotion own small energy at the same range. However; the derived spectrum is often passed through a bank of band-pass filters to better leverage the spectral distribution over the audible frequency range. Spectral features are then computed from the outputs of these filters [7].

In spite of the fact that many research works were conducted in continuous and spectral based features, recognition of emotion from acoustic attributes remains a challenging task. An example of speech emotion recognition difficulties; happiness and anger both have common acoustic traits like pitch, formant, number of times their speech crosses zero pivots. Therefore, a problem occurs during the recognition of these two sets of emotions [3].

Image processing methods can be also used with speech signal to extract the discriminated features after converting the signal into spectrogram image. Spectrogram image provides information about signal amplitudes by plotting the signal as 2D image. Wang (2014) [8] extracted texture features from speech spectrogram image. First, the spectrogram transformed as a recognizable image using Fourier transform. Cubic curve is then used to improve the contrast of spectrogram image. Next, texture features have been extracted from the spectrogram image by applying Laws' masks on the image to represent the emotion state. Finally, Support Vector Machine (SVM) is used as a classifier to obtain the results of the proposed system. Three databases were used to test the efficiency of the proposed system: Berlin Emotional Speech Database (EMO-DB), eNTERFACE corpus, and one self-recorded database. The proposed system gave accuracy ranges from 65.20% to 77.42% for the three used databases, respectively. Papakostas et al. (2017) [9] used Convolutional Neural Network (CNN) with raw spectrogram images and SVM with vector of low-level features. As it is depicted by the results of experiments, the proposed system gave an accuracy of 30% for Surrey Audio-Visual Expressed Emotion (SAVEE) database, 45% for EMOVO database and 80% for EMO-DB using SVM classifier. Hajarolasvadi and Demirel (2019) [2] used the sequence of key frames' spectrograms with a 3D-CNN for SER solution. The proposed 3D-CNN system consists of two convolutional layers for feature extraction and one fully connected layer for classification. Experiments were carried out on Ryerson Multimedia Laboratory (RML) and eNTERFACE'05 databases and the achieved results were 71.44% and 72.33%, respectively. Mustaqeem and Kwon (2020) [3] proposed a Deep Stride Convolutional Neural Network (DSCNN) model to analysis the spectrogram image of the speech signal. The locally hidden patterns in the image are discovered in convolutional layers using special strides for down-sampling the generated feature maps instead of pooling layer. After that, the extracted features are classified using fully connected layers. The proposed technique was evaluated on Interactive Emotional Dyadic Motion Capture (IEMOCAP) and Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) datasets. The system showed improvements in the achieved accuracy form the baseline results by 7.85% and 4.5%, for the two dataset, respectively. Pikramenos et al. (2020) [10] used Oriented FAST and rotated BRIEF (ORB descriptors) that were extracted from key point locations on the spectrogram image to generate an intermediate representation. First, a method similar to Bag-of-Visual-Words (BoVW) is utilized, where a visual vocabulary is built by clustering the descriptors of key points. Soft candidacy score is then used to generate the final histogram descriptors of the signal. The achieved accuracy using SVM classifier was 58.33% on SAVEE database.

Although many speech characteristics were investigated in the recognition of speech emotions, till now the researchers did not identify the best speech characteristics for this task. This is due to the similarity in the extracted features for the different emotions. The main motivation of the paper is the fact that the Extended Local Binary Pattern (ELBP) has not been reported in SER problem solution. To fill this research gap, this paper aims at finding the application ELBP in the analysis of the texture of the generated spectrogram image. The main contributions of this work can be summarized with the following points:

- ELBP descriptors are proposed to measure the efficiency of such type of features in SER Problem. ELBP gives information about the direction and amount of change in amplitude intensities for the given emotion which in effect leads to more effective feature vector.

- Deep Belief Network (DBN) classifier is used to classify the extracted descriptors and find the optimal ELBP combination based on the resulted accuracy.

The organization of the remaining paper is as follows: stages of the proposed system are demonstrated in detail in Section 2. In Section 3, the experimental analysis and evaluation of the proposed system have been conducted. Finally, the work's conclusion and ideas for future work are presented in Section 4.

2. The proposed method

The proposed system involves four important stages. First, the spectrogram image is constructed from the speech signal. After that, four different variants of ELBP are generated from the spectrogram image. Multi-block histograms are then extracted in feature extraction stage which is finally fed to DBN for emotion classification. Fig. 1 shows a general view of the proposed SER system.

2.1 Spectrogram image construction

In this stage, the spectrogram image is constructed for the given speech signal. The most commonly used approach for the construction of speech spectrogram image is by describing the amplitude of a particular frequency at a particular time with intensity in the constructed image. The following five steps are applied to build the spectrogram image for the speech signal:

1) The speech signal is first partitioned into overlapped frames with respect to the time. After that, windowing process is applied to reduce the effect of dis-connectivity at the ends of each frame.

2) Each frame is then transformed from the time domain into the frequency domain by incorporating Short Term Fourier Transform (STFT) using the following equation [11]:

$$F(v) = \frac{1}{n}\sum_{x=0}^{n-1} s(x)e^{-j2\pi\frac{(vx)}{n}} \quad (1)$$

Where, *s* is the speech frame with length *n*, *F(v)* is result of applying Fourier transform on s*(v)* and $j=\sqrt{-1}$. Based on Euler's property, Eq. (1) can be rewritten as:

$$F(v) = \frac{1}{n}\sum_{x=0}^{n-1} s(x)\left[cos\left(\frac{2\pi}{n}(vx)\right) + j\ sin\left(\frac{2\pi}{n}(vx)\right)\right] \quad (2)$$

Where, the cosine term represents the real part (*R(v)*) and the sine term represents the imaginary part (*I(v)*). The magnitude ($|F(v)|$) at a given speech signal (*v*) can be computed as:

$$Magnitude=|F(v)| = \sqrt{[R(v)]^2 + [I(v)]^2} \quad (3)$$

3) After applying STFT, a Gaussian low-pass filter is performed to drop out the high-frequency components (i.e., noise). The following transfer function is used with Gaussian low-pass filter [12]:

$$H(v) = e^{-D^2 s(v)/2D_0^2} \quad (4)$$

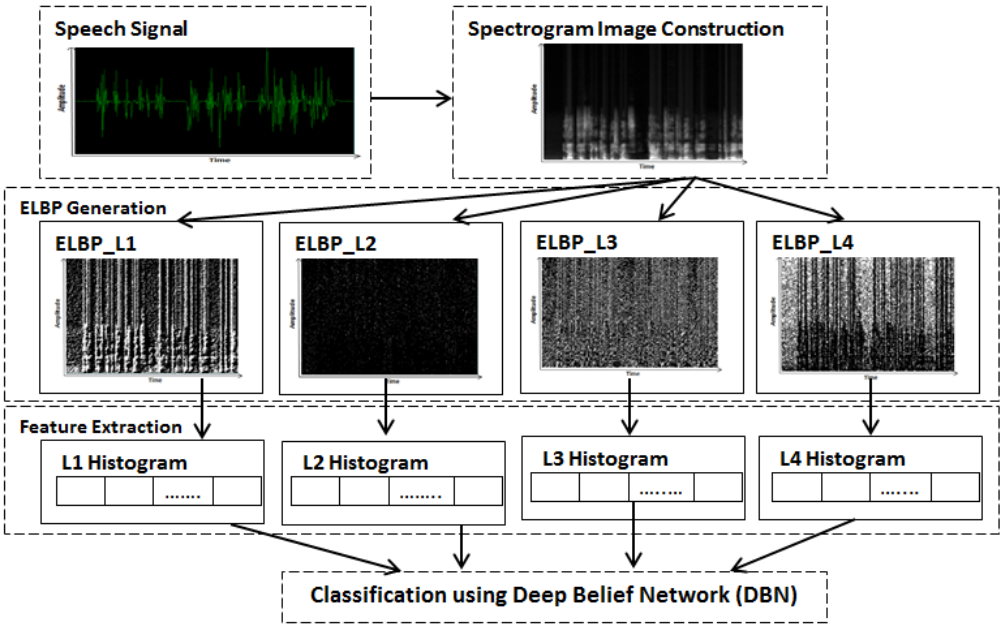


Figure. 1 General overview of the proposed SER system

Where, s is the speech frame with length n, $D(v) = [v - n/2]^2$ and D_0 represents the desired distance from the origin of STFT transform.
4) The log spectrogram is then computed, to simulate the way that human ear precepts the sound in nature (i.e., logarithmic nature). The intensity of each logged transformed value $F(v)$ within each speech frame can be calculated as following [13]:

$$Intensity\ (v) = \left(\frac{255}{max - min}\right) \times (10 * log_{10}|F(v)| - min) \quad (5)$$

Where, $Intensity\ (v)$ is the color intensity of $F(v)$, *max* and *min* represent the maximum value and minimum value in the speech frame, respectively.
5) After computing the intensity degree of each value in the frame, the frame will be represented as a column in the constructed spectrogram image.
6) Finally, to speed up ELBP generation and feature extraction tasks, the spectrogram image is down-sampled to new dimensions ($N_{wid} \times N_{hgt}$). Bilinear interpolation is utilized for this purpose with N_{wid}=512 and N_{hgt}=512.
Fig. 2 shows an example spectrogram images constructed for seven different speech signals with different emotions. As clearly depicted in the figure, the spectrogram images of the different speech signals contain a great deal of information that can help in solving SER problem.

2.2 ELBP variants generation

Spectrogram image can be viewed as a texture image. In the field of computer vision, the appearance of the texture image can be defined as a significant change in the pixel intensity and non-homogeneity between nearby pixels. To analysis the patterns of the texture image, different descriptors have been introduced by the researchers such as Scale-Invariant Feature Transform (SIFT) and Speeded Up Robust Features (SURF). Local Binary Pattern (LBP) is one of the recent advances in texture descriptors which promise a significant progress in trends of texture analysis by means of their high discriminative properties because it is illumination invariant and computational inexpensive.

The LBP demonstrates the texture by using micro-primitive components based on the statistical rules of pixels' placements. The LBP is working in a pixel-basis and can be defined as a binary code that represents the eight pixels surrounding the pixel using filter of size *fs*×*fs*. The LBP then combines all codes into a histogram, which represents features of the texture. Thus, for a filter of size 3 × 3, a histogram of size 256 will be generated.

Mathematically, the LBP for the pixel *I(x, y)* in an image (*I*) using filter of size *fs*×*fs* can be defined as follows:

$$LBP\left(I(x,y)\right) = \sum_{i=x-\left(\frac{fs}{2}\right)}^{i=x+\left(\frac{fs}{2}\right)} \sum_{j=y-\left(\frac{fs}{2}\right)}^{j=y+\left(\frac{fs}{2}\right)} s\big(I(x,y) - I(i,j)\big) \times 2^p \quad \forall\ i \neq x, j \neq y \quad (6)$$

Where, $s(v) = \begin{cases} 1 & if\ v \geq 0 \\ 0 & otherwise \end{cases}$, and p is an integer number range from 0 to 7 that represents the location of neighbour pixel in the opened window.

However; many additional information can be extracted using LBP such as change direction encoding, amount of change using extended variants. A Multi-Block Extended Version of LBP (MELBP) is proposed in this paper to highlight the important features of spectrogram image texture. Three additional variants (*L2*, *L3*, and *L4*) of LBP are generated to provide information about the amount of change in each pixel intensity with respect to its neighbours. To generate ELBP variants for the spectrogram image (*I*) using filter with size (*fs*×*fs*), the following steps are applied on each pixel (*I (x, y)*) in *I*:
1) A window of size (*fs*×*fs*) is opened around *I (x, y)*.
2) The normal LBP variant (*L1)* is first computed using Eq. (6).
3) For each pixel in the opened window (*I (i, j)*), the absolute difference (*Diff (i, j)*) between *I (x, y)* and *I (i, j)* is computed using the following equation:

$$Diff(i,j) = |I(x,y) - I(i,j)| \quad (7)$$

4) The computed difference of each pixel in the opened window is then normalized to be within the range [0-3].
5) After that, the normalized difference of each pixel is converted to the binary representation ($B_{i,j}()$).
6) An additional three variants of LBP (*L2*, *L3*, and *L4*) is computed using the following equations:

$$L2\left(I(x,y)\right) = \sum_{i=x-\left(\frac{fs}{2}\right)}^{i=x+\left(\frac{fs}{2}\right)} \sum_{j=y-\left(\frac{fs}{2}\right)}^{j=y+\left(\frac{fs}{2}\right)} B_{i,j}\,(0) \quad \forall\ i \neq x, j \neq y \quad (8)$$

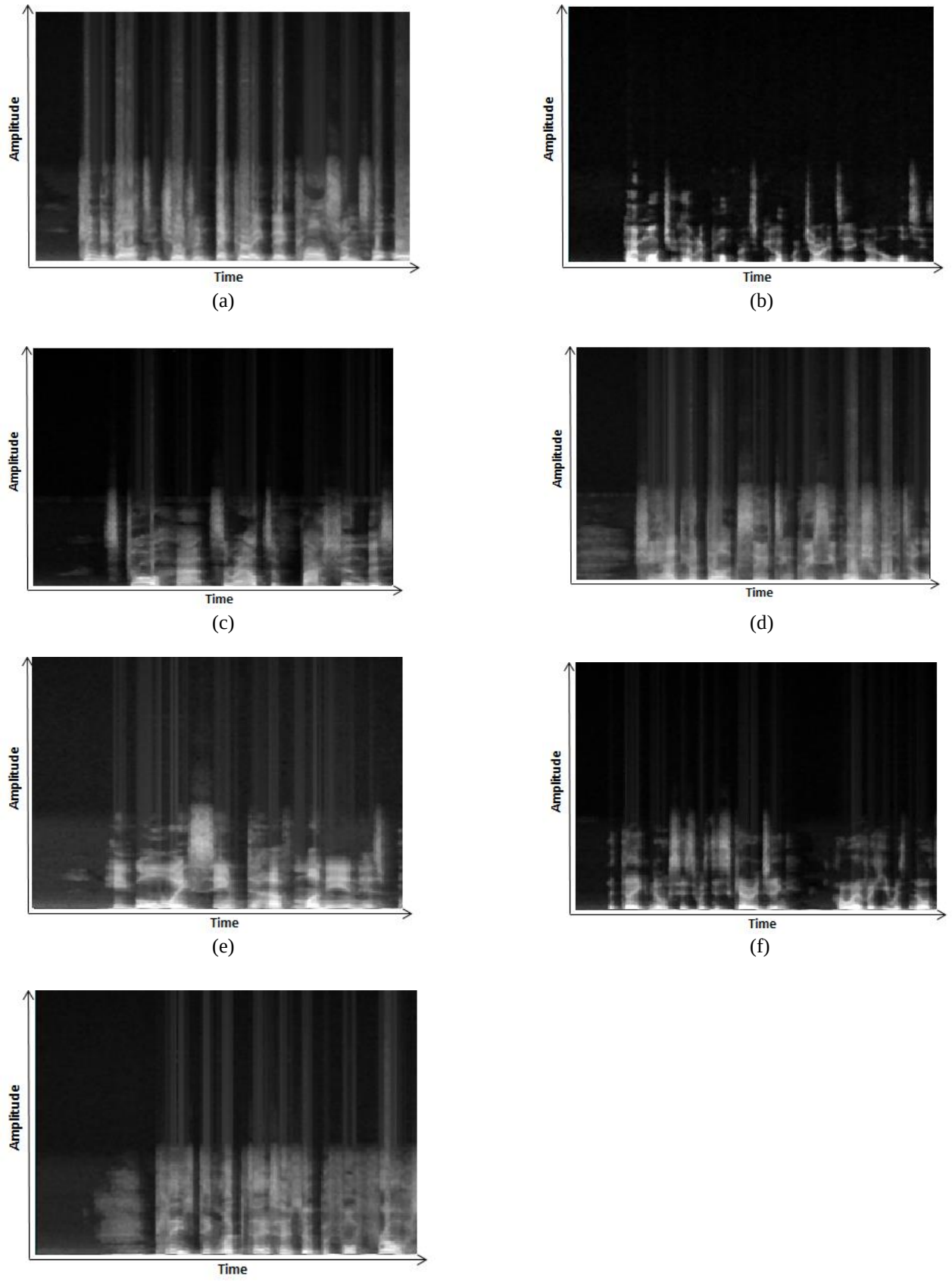


Figure. 2 Spectrogram images for seven different emotions: (a) anger emotion, (b) disgust emotion, (c) fear emotion, (d) happy emotion, (e) natural emotion ,(f) sad emotion, and (g) surprise emotion

$$L3\left(I(x,y)\right) = \sum_{i=x-\left(\frac{fs}{2}\right)}^{i=x+\left(\frac{fs}{2}\right)} \sum_{j=y-\left(\frac{fs}{2}\right)}^{j=y+\left(\frac{fs}{2}\right)} \left(B_{i,j}\,(1)\ \times 2\right) \quad \forall\ i \neq x\ , j \neq y \tag{9}$$

$$L4\left(I(x,y)\right) = \sum_{i=x-\left(\frac{fs}{2}\right)}^{i=x+\left(\frac{fs}{2}\right)} \sum_{j=y-\left(\frac{fs}{2}\right)}^{j=y+\left(\frac{fs}{2}\right)} \left(B_{i,j}\,(2)\ \times 4\right) \quad \forall\ i \neq x\ , j \neq y \tag{10}$$

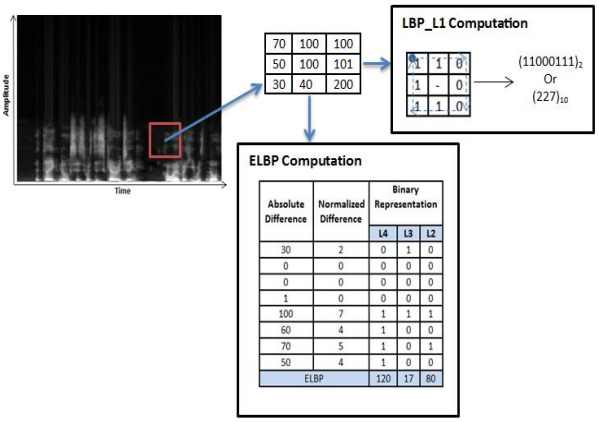


Figure. 3 A numerical example demonstrates the idea of ELBP

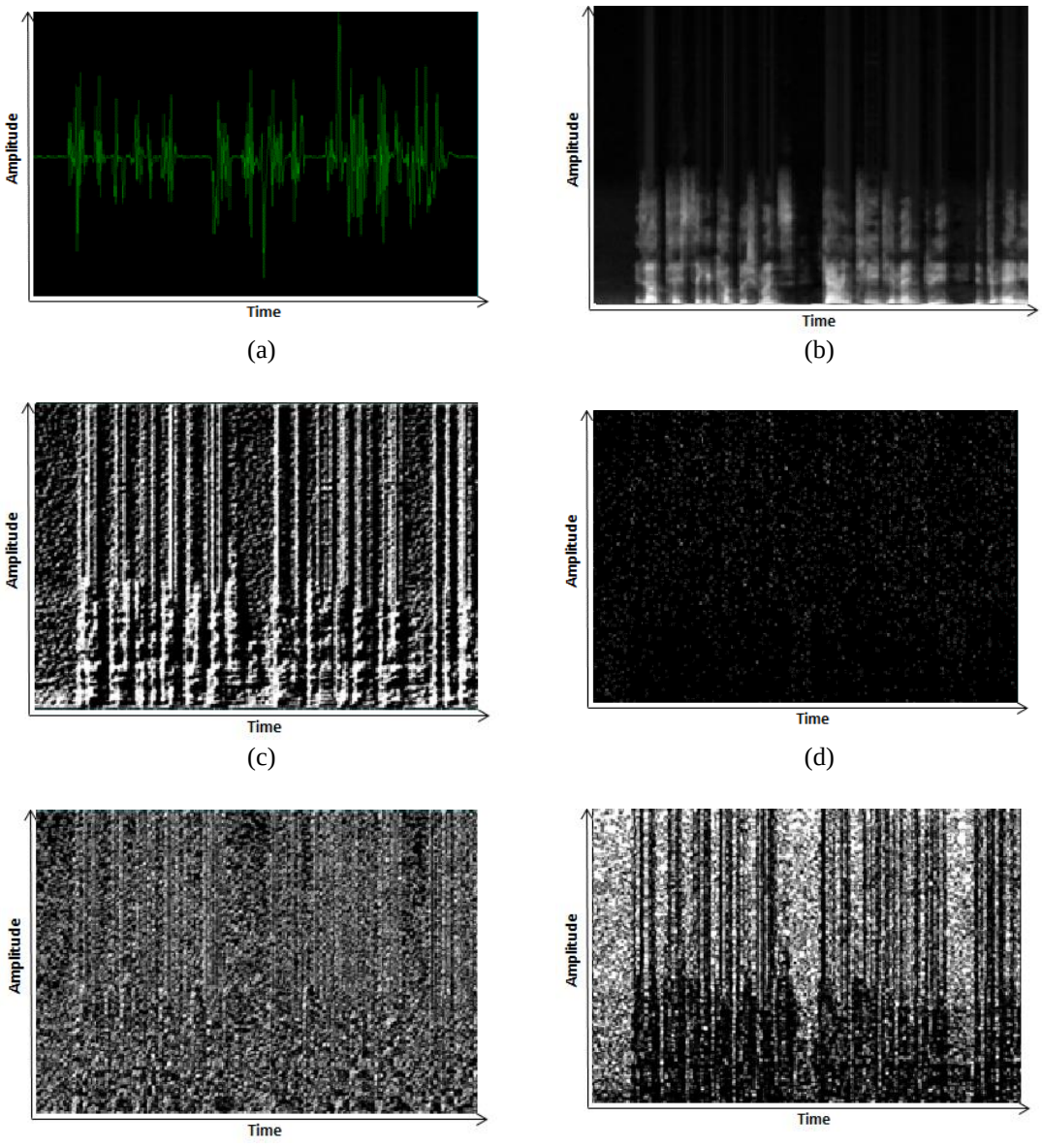


Figure. 4 Results of applying ELBP on an example speech signal: (a) speech signal, (b) spectrogram image, (c) L1, (d) L2, (e) L3, and (f) L4

Where, fs is filter size, $B_{i,j}$ is the binary representation of the difference between the current pixel $I(x,y)$ and its neighbour pixel $I(i,j)$. *L2* gives information about the amount of the small change in pixel intensity with respect to its neighbours (i.e., the first Least Significant Bits (LSB) with weight equals to $2^0 = 1$) while *L3* gives information about the amount of change in the second LSB (with weight equals to$2^1 = 2$) and *L4* gives information about the amount of change in the most significant bit (with weight equals to$2^2 = 4$). Fig. 3 shows a numerical example of applying ELBP on an example spectrogram image while Fig. 4 demonstrates graphical results of ELBP variants when applied on a sample spectrogram image.

2.3 Feature extraction

Four vectors of attributes are generated from the results of applying ELBP. Each vector represents the histogram of each ELBP variant. Feature extraction stage can be summarized with the following three steps:

1) Each intensity value of ELBP variants (i.e., *L1*, *L2*, *L3*, *L4*) are first normalized to be within the range [0-3] to avoid building very large feature vector.

2) For each normalized variant of ELBP, the image is divided into blocks of size (*Bs*×*Bs*) and the histogram of each block is then computed.

3) The histograms of all blocks are then merged together to form the final feature vector that represents the Multi-block Extended Local Binary (MELBP) of that ELBP variant.

The final result of applying MELBP on the spectrogram image is four feature vectors, each with size equals to the number of blocks × 4 (histogram size). Fig. 5 demonstrates the main idea of feature extraction stage.

2.4 Classification

The Deep Belief Network (DBN) classifier is utilized in the proposed SER system due to its strong classification ability. DBN consists of stacked Restricted Boltzmann Machines (RBMs) that firstly performs the unsupervised learning to avoid vanishing problem that occurs when training the network from scratch weights (i.e., random weights). The weights of the network are then fine-tuned using supervised learning (i.e., regular back propagation training). RBM is a Boltzmann machine which has multiple hidden layers and one visible layer as shown Fig. 6 [14]. The classification stage involves two steps which are training step and testing step. In training step, the DBN will be trained with the

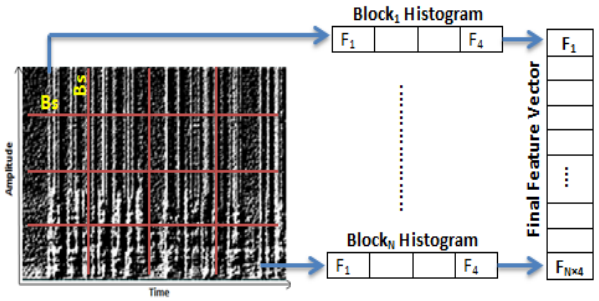


Figure. 5 The main idea of feature extraction stage

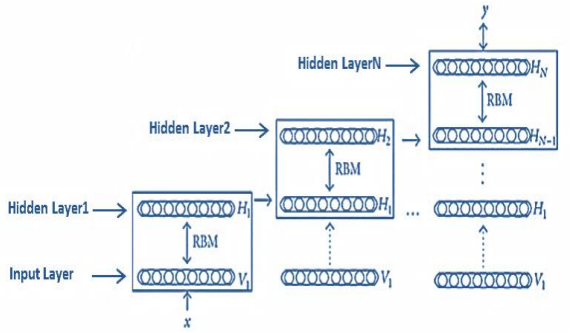


Figure. 6 Architecture of Deep Belief Network (DBN)

training part of the database to find the optimal DBN architecture and adjust the network weights. The resulted weights are then fed to testing step for guessing the emotion of the given MELBP feature vector.

3. Experimental results and analysis

3.1 Database

The proposed SER system is evaluated using samples taken from the Surrey Audio-Visual Expressed Emotion (SAVEE) database [15]. The SAVEE database was established at the University of Surrey in the United Kingdom. It includes recordings for six basic emotions, i.e. anger (Ang), disgust (Dis), fear (Fea), happiness (Hap), sadness (Sad), and surprise (Sur), as well as natural (Nat) state. The database contains 480 native British utterances formed of 60 samples for each of the emotional states except for natural emotion which includes 120 samples. Emotional and text prompts were displayed in front of the subject on a monitor while each video is recoding. In each prompt, a video clip and three images were included such that the text was divided into three groups for each emotion, in order to avoid fatigue. The length of the recorded video ranged from one second to seven seconds. The samples were captured with a video camera sampled at 60 frames per second (fps) with resolution of about 81,920 pixels and mono voice signals in a 16-bit format sampled at 44,000 Hz. The captured videos were

evaluated by ten subjects, under audio, visual and audio-visual conditions

For system evaluation purpose, a total of 420 samples are used (i.e., 60 samples per each emotion) to make sure a balance in the number of samples within each class. To avoid over fitting during training process, 80% of samples within each emotion are used for system training and 20 % for testing purpose.

3.2 Results

Different experiments are conducted to find the optimal configuration of the proposed SER parameters. It is found that the best length for speech frame is 512 samples with overlap equals to 50%. While the optimal D_0 value in a Gaussian low-pass filter is 100. The ELBP filter size (*fs*) is 3. On the other hand; experiments are performed to study the effect of changing the value of Bs parameter on the final accuracy of the proposed SER system. The measure that is used in system evaluation is the classification accuracy which can calculate as follows [16,17]:

$$Accuracy = \frac{correct}{Total} \times 100 \qquad (11)$$

Where correct represents the number of speech signals that correctly classified and Total is the total number of speech samples in the database.

Table 1 shows results that are obtained using *L1*, *L2*, *L3*, and *L4* as a source for feature extraction task, individually, using different values of Bs. As shown in the table, the best achieved accuracy is 45.95% for *L1*, 51.19% for *L2*, 54.76% for *L3* and 46.67% when Bs=64. *L3* gives the highest accuracy with DBN configuration equals to number of hidden nodes (H) =16 and learn rate (LR) =0.8. This refers to the fact that *L3* variant of ELBP holds most of discriminative information about the emotion state of the spectrogram image. Table 2 shows the confusion matrix for speech samples distribution within each emotion for the highest accuracy achieved in Table 1.

As shown in Table 2, happiness is the most recognized emotion with accuracy equals to 70% while sad emotion is the least recognized one with accuracy of 41.67%. Most of the overlap occurs between fear and sad emotions which refers to the similarity in the traits of spectrogram representation of these two emotions.

The combinations of different ELBP variants are also tried to assess the accuracy that can be achieved for each possible combination. Table 3 shows results that are achieved using different forms of ELBP combinations as a source for feature extraction task

Table 1. Accuracy (%) achieved using each of MELBP variants, individually

ELBP	Bs=8	Bs=16	Bs=32	Bs=64	Bs=128
L1	16.67	29.52	35.71	45.95	30.00
L2	19.05	28.10	42.86	51.19	29.76
L3	21.43	28.10	45.24	**54.76**	30.00
L4	18.57	29.05	41.67	46.67	29.76

Table 2. Confusion matrix (%) for the best result achieved in Table 1

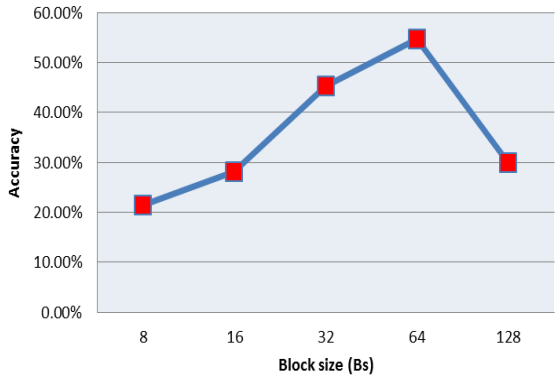
	Ang	Dis	Fea	Nat	Hap	Sad	Sur
Ang	**66.6**	1.67	1.67	1.67	13.33	3.33	11.67
Dis	5.00	**50.0**	13.33	10.00	3.33	16.67	1.67
Fea	1.67	5.00	**51.67**	8.33	1.67	28.33	3.33
Nat	3.33	13.33	16.67	**45.00**	5.00	13.33	3.33
Hap	8.33	0.00	1.67	6.67	**70.00**	0.00	13.33
Sad	0.00	8.33	33.33	16.67	0.00	**41.67**	0.00
Sur	11.67	1.67	0.00	10.00	16.67	1.67	**58.33**

Table 3. Accuracy (%) achieved using different combinations of MELBP variants

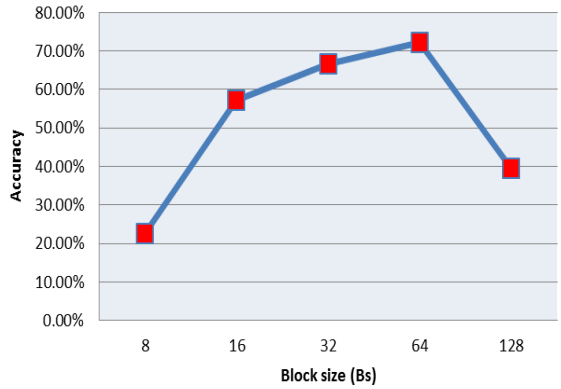
ELBP	Bs=8	Bs=16	Bs=32	Bs=64	Bs=128
L1+L2	14.29	19.05	42.86	62.86	39.52
L1+L3	20.71	47.86	59.52	71.19	39.52
L1+L4	22.62	57.14	66.67	**72.14**	39.52
L2+L3	15.48	21.90	35.71	48.57	34.76
L2+L4	14.29	20.23	28.33	45.95	32.38
L3+L4	15.48	18.57	31.67	33.57	31.43
L1+L2+L3	20.23	44.76	47.62	61.67	38.80
L1+L2+L4	14.29	15.71	37.62	47.62	34.29
L2+L3+L4	14.29	14.52	26.67	41.67	32.86

Table 4. Confusion matrix (%) for the best result achieved in Table 3

	Ang	Dis	Fea	Nat	Hap	Sad	Sur
Ang	**88.33**	0.00	0.00	1.67	5.00	0.00	5.00
Dis	0.00	**71.67**	8.33	11.67	0.00	8.33	0.00
Fea	0.00	13.33	**60.00**	3.33	0.00	23.33	0.00
Nat	6.67	5.00	11.67	**55.00**	5.00	11.67	5.00
Hap	8.33	0.00	0.00	3.33	**85.00**	0.00	3.33
Sad	3.33	5.00	8.33	16.67	1.67	**61.67**	3.33
Sur	5.00	0.00	0.00	6.67	5.00	0.00	**83.33**



(a)



(b)

Figure. 7 The relationship between block size and accuracy: (a) using ELBP variants individually and (b) using combined ELBP variants

Table 5. Comparison between the proposed SER system and other works utilized spectrogram image and SAVEE database

Authors	Method	Database	Achieved accuracy
Papakostas et al. (2017) [9]	Spectrogram image, CNN and SVM	SAVEE	30.00%
Pikramenos et al. (2020) [10]	Spectrogram image with Oriented FAST, rotated BRIEF and SVM	SAVEE	58.33%
Proposed SER system	**Spectrogram image, MELBP and DBN**	SAVEE	**72.14%**

with different values for Bs. As shown in the table, the best accuracy (72.14%) has been achieved when combining feature vectors of *L1* and *L4* variants with Bs=64, H=20 and LR=0.2. The confusion matrix for the distribution of speech samples in the best combination is shown in Table 4.

As shown in Table 4, Anger is the most recognized emotion with accuracy equals to 88.33% and natural is the least recognized one with accuracy equals to 55.00%. The overlap between fear and sad emotions is reduced when using combined feature vector. However; some overlap still exists between natural and remaining emotions due to the similarly in some features of natural state and other emotions.

3.3 Results analysis

As shown in results section, when each of ELBP variants is used separately, low accuracy was achieved within the four different variants. On the other hand, the combination of ELBP descriptors increases the accuracy with an about 17.38% from the best accuracy that was achieved when ELBP descriptors used in separate fashion. This refers to the fact that weak MELP features can lead to strong feature vector when they are combined together.

On the other hand, as Bs increases the accuracy also increases as shown in Fig. 7. The main reason behinds that is the size of the resulted feature vector will be reduced which in effect enables the DBN classifier to interpret the patterns involved within these features. DBN efficiency decreases when it fed with very long feature vector because it becomes difficult to find the common patterns among a large number of attributes. However; when the Bs reached the value 128, the accuracy is dropped out which refers to the loss in minute detail (i.e., the local description of spectrogram image) in the resulted feature vector because of the exaggerated block size.

3.4 Comparison with other works

Table 5 shows a comparison made between the proposed SER system and other works which are worked on spectrogram representation of speech signal and utilized the same database (i.e., SAVEE database). The table also shows the used methods and achieved accuracy by each work. As it is clearly shown in the table, the accuracy that are achieved by the proposed SER system outperforms those are achieved by [9] and [10] with an improvement equals to 13.81%.

4. Conclusion

A speech emotion recognition system has been presented in this paper based on the spectrogram representation of speech signals. However; instead of analysis the constructed spectrogram image directly, the extended LBP representations of the image are generated. The ELBP helps in highlighting the micro patterns in the texture of spectrogram image and in

effect leads to more effective features. Deep belief network also gives the proposed SER system the power in understanding the hidden patterns within the generated features that are extracted from MELBP variants. Experimental results showed that an accuracy of about 72.14% was achieved on SAVEE database using feature vector made up of combination from MELBP features. As a future work, deep neural network can be used with ELBP where the network can be fed with the different spectrogram images represented by ELBP variants.

Conflicts of Interest

Suhaila N. Mohammed and Alia K. Abdul Hassan declare that they have no conflict of interest.

Author Contributions

Conceptualization, methodology and implementation, writing—original draft preparation, Suhaila N. Mohammed; writing—review, editing, supervision and funding acquisition, Alia K. Abdul Hassan.

References

[1] P. Barros and S. Wermter, "Developing Crossmodal Expression Recognition Based on A Deep Neural Model", *Adaptive Behavior*, Vol. 24, No. 5, pp. 373–396, 2016.

[2] N. Hajarolasvadi and H. Demirel, "3D CNN-Based Speech Emotion Recognition Using K-Means Clustering and Spectrograms", *Entropy*, Vol. 21, No. 479, pp. 1-17, 2019.

[3] M. Mustaqeem and S. Kwon, "A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition", *Sensors*, Vol. 20, No. 183, pp. 1-15, 2020.

[4] S. Sondhi, M. Khan, R. Vijay, A. Salhan, and S. Chouhan, "Acoustic Analysis of Speech under Stress", *International Journal of Bioinformatics Research and Applications*, Vol. 11, No. 5, pp. 417-432, 2015.

[5] K. Wang, N. An, B. Li, Y. Zhang, and L. Li, "Speech Emotion Recognition Using Fourier Parameters", *IEEE Transactions on Affective Computing*, Vol. 6, No. 1, pp. 69-75, 2015.

[6] D. Bitouk, R. Verma, and A. Nenkova, "Class-Level Spectral Features for Emotion Recognition", *Speech Communication*, Vol. 52, pp. 613–625, 2010.

[7] S. Mohammed and A. Abdul Hassan, "A Survey on Emotion Recognition for Human Robot Interaction", *Journal of Computing and Information Technology*, Vol. 27, No. 4, pp. 47-68, 2019.

[8] K. Wang, "The Feature Extraction Based on Texture Image Information for Emotion Sensing in Speech", *Sensors*, Vol. 14, No. 9, pp. 16692-16714, 2014.

[9] M. Papakostas, G. Siantikos, T. Giannakopoulos, E. Spyrou, and D. Sgouropoulos, "Recognizing Emotional States Using Speech Information", *Advances in Experimental Medicine and Biology*, Vol. 989, pp.155-164, 2017.

[10] G. Pikramenos, G. Smyrnis, I. Vernikos, T. Konidaris, E. Spyrou, and S. Perantonis, "Sentiment Analysis from Sound Spectrograms via Soft BoVW and Temporal Structure Modelling", In: *Proc. of the 9th International Conf. on Pattern Recognition Applications and Methods*, Valletta, Malta, pp. 361–369, 2020.

[11] S. Mohammed, A. Jabir, and Z. Abbas, "Spin-Image Descriptors for Text-Independent Speaker Recognition", In: *Proc. of Saeed F., Mohammed F., Gazem N. (eds) Emerging Trends in Intelligent Computing and Informatics.* IRICT 2019. *Advances in Intelligent Systems and Computing*, Vol. 1073, Springer, Cham, 2019.

[12] A. Makandar and B. Halalli, "Image Enhancement Techniques Using Highpass and Lowpass Filters", *International Journal of Computer Applications*, Vol. 109, No. 14, pp. 12-15, 2015.

[13] F. Alías, J. Socoro, and X. Sevillano, "A Review of Physical and Perceptual Feature Extraction Techniques for Speech, Music and Environmental Sound", *Applied Sciences*, Vol. 6, No. 143, pp. 1-44, 2016.

[14] A. Bondarenko and A. Borisov, "Research on the Classification Ability of Deep Belief Networks on Small and Medium Datasets", *Information Technology and Management Science*, Vol. 16, pp. 60-65, 2013.

[15] Surrey Audio-Visual Expressed Emotion (SAVEE) Database Website, http://kahlan.eps.surrey.ac.uk/savee/

[16] A. Hassan and S. Mohammed, "A Novel Facial Emotion Recognition Scheme Based on Graph Mining", *Defence Technology*, 2019. http://dx.doi.org/10.1016/j.dt.2019.12.006

[17] S. Mohammed, F. Alkinani, and Y. Hassan, "Automatic Computer Aided Diagnostic for COVID-19 Based on Chest X-Ray Image and Particle Swarm Intelligence", *International Journal of Intelligent Engineering and Systems*, Vol.13, No.5, pp. 63–73, 2020.