Development of Speech Emotion Recognition System based on Discrete Wavelet Transform (DWT) and Voice Segmentation

Hertog Nugroho[1] and Raya Nadlira Nurul Fuadiyah[2]

[1]Electric Engineering Department, Politeknik Negeri Bandung
Jalan Gegerkalong Hilir, Ciwaruga, Bandung, Indonesia, 40559
[2]State High School 3 Bandung
Jalan Belitung No. 8, Bandung, West Java, 40113
[1]hertog@polban.ac.id, [2]raya.nadliranf27@sman3bandung.sch.id

Abstract: Speech emotion recognition has been investigated widely in the field of human-machine interaction. Efforts to develop recognition systems have been reported. Some of them focused on features, and others focused on classifiers. However, most of them segmented the speech signal into fixed-length frames and took the features from them. The scheme contradicts the physiological and psychological studies that emotion information is contained in long continuous voice parts of the speech signals. This study proposes a combination of Discrete Wavelet Transform (DWT) decomposition, voice part segmentation, a scheme to determine a fixed number of segments, and prosodic and spectral features to build an emotion recognition system. The voice part segmentation is adopted to accommodate the above studies, and the DWT decomposition allows the selection of the best system performance. The system has been validated with the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) dataset at several DWT decomposition levels, using prosodic features (zero-crossing rate, energy, peak), spectral features (Fourier coefficients, cepstrum), and a Multi-Layer Perceptron Neural Network as a classifier. The result shows that the best performance of this system to classify eight categories of emotion (neutral, calm, happy, sad, angry, fear, disgust, and surprised) is 98% accuracy on level 6 of DWT decomposition.

Keywords: speech emotion recognition; Discrete Wavelet Transform; voice segmentation; feature extraction; Multi-Layer Perceptron;

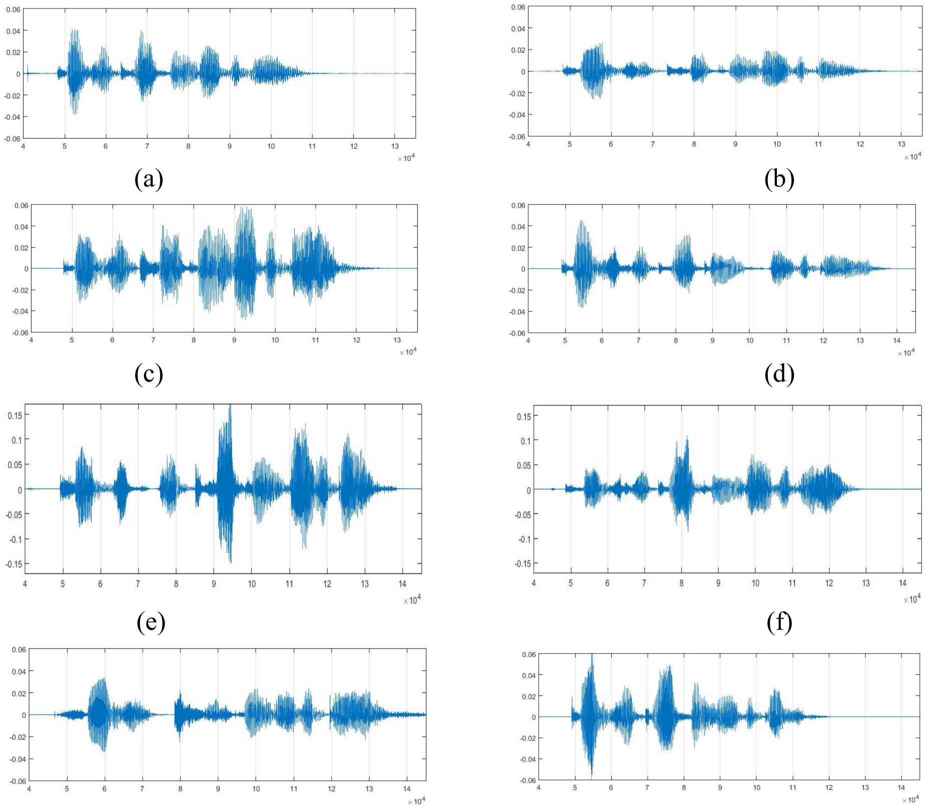
1. Introduction

Over the last decade, the interest in speech emotion recognition has increased in speech and language processing [1], [2]. Speech Emotion Recognition (SER) is particularly useful in specific applications such as tutoring systems that detect the learner's state, call-center systems for detecting consumer's states, and enhancing the service quality. Although recognition of emotions has benefits, there are still problems in developing a method to recognize emotion from the speech signal because the speaking styles of the speakers are different from person to person [3], [4]. Physiological studies show that expressing emotion has a beginning, a rising side, a peak, and a falling side [5], as shown in Figure 1. The speech segment should be long enough to capture the possible information.

Many works of Speech Emotion Recognition have been reported using fixed-length speech segments. Xianxin Ke *et al.* [6] use 33-dimensional feature parameters derived from energy characteristics, fundamental frequency characteristics, zero-crossing rate (ZCR), Mel Frequency Cepstrum Coefficient (MFCC), and formant from each segment, and use Continuous Hidden Markov Model (CHMM) to classify five emotional states: happiness, anger, sadness, fear, and calm. They use the Berlin Emotional Speech Library dataset [7] for experiments, conduct several configurations, and achieve the best 67.83% average accuracy with Principle Component Analysis feature reduction using 33-D features. Gao *et al.* [8] propose signal segmentation using Depth First Search (DFS) algorithm to decide the segment duration and overlap, extracted pitch, MFCC, Line Spectral Pairs (LSP), intensity, and ZCR from each segment (which they called

Received: January 22nd, 2022. Accepted: September 19th, 2022
DOI: 10.15676/ijeei.2022.14.3.7

'local features'), perform smoothing and normalization on them and then compute global features using Open-Source Media Interpretation by Large feature-space Extraction (Open SMILE) toolkit [9]. They adopt a linear kernel SVM with minimal sequential optimization (SMO) for the classifier to classify seven emotions (angry, boredom, disgust, fear, happy, neutral, and sad). They validate their method with the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) [10] and Emotional Speech Database (EMODB) [7] datasets and achieve 79.4% and 87.3% accuracies, respectively.



(a) (b)

(c) (d)

(e) (f)

(g) (h)

Figure 1. Examples of speech signals uttered by one actor speaking a sentence ("Kids are talking by the door") with eight classes of emotion: (a) neutral, (b) calm, (c) happy, (d) sad, (e) angry, (f) fearful, (g) disgust, and (h) surprised. Courtesy of the RAVDESS dataset [10].

Several works adopt currently popular deep learning models to improve emotion classification. For example, Mustaqeem *et al.* [11] propose Deep Stride Convolutional Neural Network (CNN) to classify speakers' emotions. To accommodate 2-D input for the classifier, they convert 1-D audio signals into a 2-D spectrogram through STFT. Before conversion, they clean the audio signals to remove the background noises, silent portions, and other irrelevant information from speech signals using adaptive threshold-based preprocessing. Finally, they validate their method with the RAVDESS dataset with four classes of emotions and the Interactive emotional dyadic motion capture (IEMOCAP) dataset [12] with eight classes of emotions and achieve 79.5% and 81.75% accuracies, respectively. Furthermore, Mustaqeem and other authors [13] attempt to improve [11] by proposing Radial Basis Function Network (RBFN) similarity measurement to select audio segments in clusters, convert the segments into spectrograms, and feed the spectrograms to the Recurrent Neural Network (RNN), which is a combination of Resnet-101 Architecture [14] and Bidirectional Long Short-Term Memory

(BiLSTM) for the classifier. Their performance achieves 72.25%, 85.57%, and 77.02% accuracy for IEMOCAP, EMODB, and RAVDESS, respectively.

Spectrogram-based audio information is also used by Zeng *et al.* [15]. They adopt Multi-Task Learning (MTL) idea and propose a variation of a deep Residual Networks (ResNets) model with the addition of a gate mechanism, which they call Gated Residual Neural Networks (GResNets). They claim that the MTL approach can have better performance than task-specific models. They attempt the various configurations of the model with the RAVDESS dataset to classify eight categories of emotion and achieve the best performance of 64.48% accuracy.

The MTL approach is also adopted by Biqiao *et al.* [16]. They propose three shared models to classify emotions from speech and song inputs simultaneously. Their performance in song inputs achieved the highest accuracy of 99% for RAVDESS and 98% for the University of Michigan Song and Speech Emotion Dataset (UMSSED) [17]. In a speech category, the highest performance was achieved at 98.19% accuracy for RAVDESS using spectral features and at 95.83% accuracy for UMSSED using Energy and MFCC features.

All the above works divide input signals into fixed-length segments. It is necessary since the classifier needs fixed-size inputs, and it is easy to extract prosodic and spectral features from fixed-length segments. However, it is well known that emotion information is contained in voice parts of audio signals [18]. Therefore, it is not strategic to fragment the signals into fixed-length segments without knowing the start-end position of the emotion information parts. This idea is exploited exclusively in [19], where the authors conducted experiments to compare frame-based (fixed-length) segments with voiced-based (variable-length) segments on emotion recognition and demonstrated that the voiced-based approach performed better than the frame-based one. The conclusion motivates us to adopt the approach.

Another issue establishing the foundation of our method is the adoption of wavelets. Discrete time-series signal analysis can be done by either Discrete Fourier Transform (DFT) or Discrete Wavelet Transform (DWT), among others. The Fourier-transformed signal gives a frequency distribution of the original time-domain signal. However, the characteristics of the transformed signal cannot be used to analyze a dynamically changed signal such as speech. The Short-Time Fourier Transform (STFT) is an effort to overcome this limitation by segmenting the signal into several frames with a fixed-length window, adding the time information of the signal. This idea is adopted by [11], [13], and [15] to convert audio signals into spectrograms. However, since STFT uses fixed-length window size, it still lacks flexibility. On the other hand, the DWT provides flexibility in window size as a function of analyzing frequency. Furthermore, the analysis function of DWT can be selected with more freedom. A deep study between DFT and DWT can be found in [20].

Efforts to recognize speech emotion using wavelets have been reported. In [21], they combine features from Continuous Morlet Wavelet Transform, prosodic features (LPC, energy, ZCR, entropy), and statistic features (maximum, minimum, mean) to form feature vectors. The process includes PCA feature reduction, Non-Negative Matrix Factorization, and various SVM classification methods. For the experiment, they apply the RAVDESS dataset with eight emotion categories. Silent areas at the beginning of audio signals are removed manually. They achieve the best performance of 60.1% accuracy using a Quadratic SVM classifier. A similar approach is also reported in [22]. They derive Linear Predictive Cepstral Coefficients (LPCC) and MFCC features from wavelet coefficients which they call WLPCC and WMFCC, respectively. In an experiment, they applied EMODB and Surrey Audio-Visual Expressed Emotion (SAVEE) [23] datasets and attempted several combinations of features to classify five emotion categories using the Radial Basis Function Network (RBFN) classifier. The best performance of 93.67% average accuracy is achieved on a combination of WLPCC and WMFCC with Vector Quantization feature reduction.

2. The Proposed Method

Similar to the above works, our objective is to build Speech Emotion Recognition with high accuracy. Our method starts with removing silent parts at the beginning and end of speech signals. The idea is similar to [11] but uses a different method, which is explained in the next section. Then, wavelet decomposition through low-frequency filters is performed until the allowable level, determined by Nyquist-Shannon Theorem [24]. Unlike [21] and [22], where the wavelet was adopted to build features taken from all decomposition levels, we use wavelet to achieve several levels of audio signals and perform feature extraction and classifications on each level. We hypothesize that the most representative information, including emotion, is not contained at all levels but at a certain level. Next, the signals are segmented into voice and unvoice parts. The basic idea is similar to [19] but different in implementation. First, they used different numbers of segments for each class of emotion, which were determined manually. Then, they conducted an experiment on each class using ten folds cross-validation scheme and a simple linear classifier. Meanwhile, we have developed a scheme to determine the same number of voice segments for all classes, extract features, and conduct experiments on all classes simultaneously so that the system can run without human intervention. Our proposed scheme is illustrated in Figure 2.

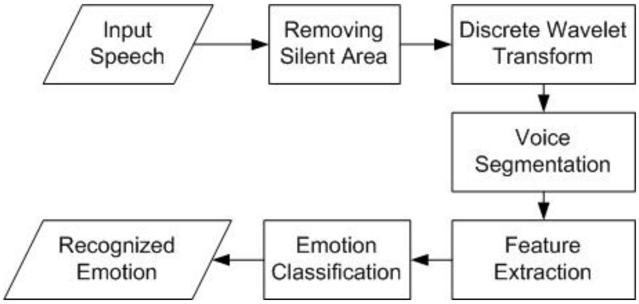


Figure 2. Block diagram of our proposed speech emotion recognition system.

The contribution of this work can be summarized as follows:

a. Wavelet decomposition is adopted to obtain filtered signals at several allowable levels, extract features at each level, and then perform classification at each level. The level providing the best performance is selected to be included in our proposed system (explained in subsection *B*).
b. Since the classifier requires the same length of feature vectors, we developed a scheme to build fixed-length feature vectors automatically, even though the audio signals are varied in length and volume (explained in subsection *D*).

The following sections describe the process of removing silent area, decomposing DWT, segmenting voice part, determining the number of voice segments, extracting features, and classifying emotion.

A. Removing Silent Area

The silent areas usually appear at the beginning and end of the speech signals as well as among spoken sentences. If the silent parts of the signal are kept, they will degrade the system performance [3]. Therefore, only the signal parts containing the actual speech are needed for speech emotion recognition.

This study applies two-step algorithms to remove the silent area (See Algorithms 1 & 2). First, the *mean* and *standard deviation* of the speech signals are calculated and used as a threshold. From statistical literature, assuming a normal distribution, the mean value represents the average value of the signal, and also assuming high S/N ratio, the values of the noise signal would be far below the mean value. So the threshold needs to be somewhere below the mean value. Meanwhile, standard deviation measures the dispersion of the signal, and the signal can be said to be distributed 'inside' the standard deviation range (not actually, since there would be outliers). Therefore, we set the threshold (*th*) equals (*α*/*standard deviation*)**mean*, while *α*,

$0<\alpha<1$ represents our confidence that the noise signals would be below *th*. In this study, α is decided from observation.

Algorithm 1. Identify voice and un-voice segments in speech signal

```
Input: speech signal y, length of signal L
Output : state of samples in signal state
calculate meanVal equals mean of y
calculate StDev equals standard deviation of y
Let th equals $\alpha * meanVal / StDev$
for i is 1 to L
    if abs(y(i)) > th then
        state(i) to 1
    else
        state(i) to 0
    end if
end for
```

Algorithm 2. Identify voice frame & discard frames with unvoice state

```
Input: samples per frame w and length of y signal L
Output : voice signal VoicedSignal
usefulSamples = L – modulus(L, w)
FrameCount = usefulSamples/w
voicedFrameCount = 0
for i is 1 to FrameCount
    let cVoiced and cUnvoiced is 0
    for j=(i-1)*w+1 to (i*w)
        if state(j) is 1 then
            cVoiced ++
        else
            cUnvoiced ++
        endif
    end for

    if cVoiced > cUnvoiced then
        voicedFrameCount ++
        voicedState(i)=1
    else
        voicedState(i)=0
    end if
end for
Let k equals 0
for i is 1 to frameCount
    if voicedState(i) is 1 then
        for j=(i-1)*w+1 to i*w
            VoicedSignal(k) = y(i)
            k ++
        end for
    end if
end for
```

Next, the signal is segmented into fixed-length frames, and the states of each frame are determined, whether it is a voice frame (above *th*) or an un-voice one. The risk of using fixed-length frames is that the parts of the voice signals can accidentally be removed at the beginning and the end of the signals. However, it is still insignificant compared to the rest of the speech signal areas. Figure 3 illustrates the effectiveness of our algorithms.

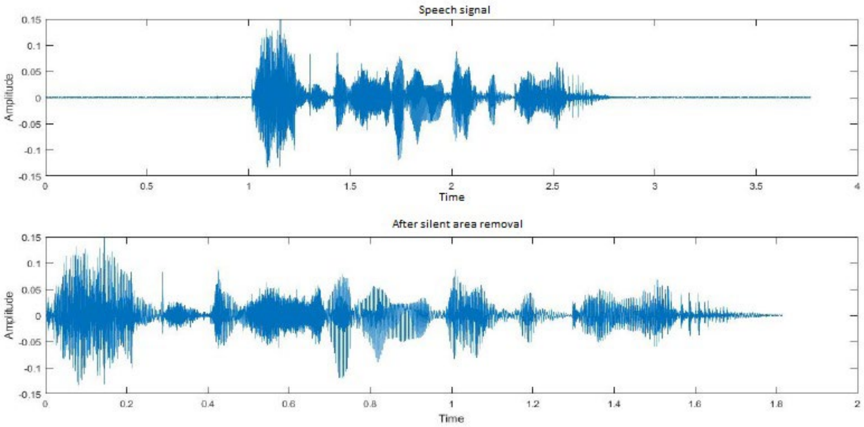


Figure 3. Before (top) and after (bottom) results of removing a silent area.

B. Determining the Level of DWT Signal

As mentioned, DWT is implemented to decompose the speech signal to several levels, and this method is tested on each level. Figure 3 illustrates the decomposition of 2-level DWT using a filter bank structure. A signal $x[n]$ is decomposed into two signals through LPF (low-frequency part) and HPF (high-frequency part), followed by down-sampling (by a factor of 2). Figure 4 constitutes a one-level decomposition. Since this study is interested in low-frequency components, further decomposition is done on the low-frequency part.

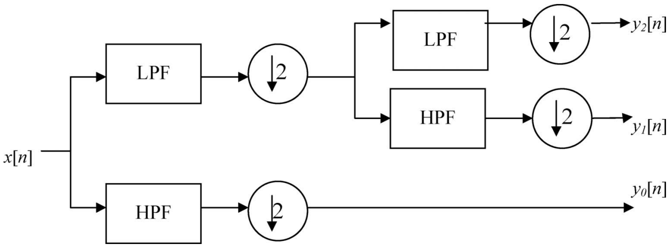


Figure 4. A DWT implementation using a filter bank structure (picture courtesy of [25]).

Since this experiment is working on human speech signals, the level of DWT signal decomposition is limited by human speech frequency (85-180 Hz for adult males and 165-255 Hz for adult females [26]) and the sampling frequency (*f*s). Therefore, the maximum female speaker frequency *f*max (255 Hz) is selected. Eq. (1) determines the allowable level of DWT (*DWT*).

$$l_{DWT} = \lfloor \log_2(f_s/(2 * f_{max})) \rfloor \qquad (1)$$

where a value of 2 is from Nyquist-Shannon Theorem [24]. Implementing Eq. (1), and setting f_s = 48k bps, and $f_{\max}$ = 255 Hz, l_{DWT} can be calculated equals to 6. It means that the DWT decomposition for the above case can go from level 1 to level 6.

Decomposition level affects the performance of recognition. Since at each level of decomposition, the signal was downsampled through low-frequency part, the high-frequency part of the signal (including noise) is removed. Further decomposition will reduce the noise even further. Therefore, a signal decomposition from level 4 to level 6 is selected because it provided better signal information and better noise reduction. In this study, *db4* is used for wavelet scaling function [27], [28].

C. Segmenting Voice Parts

As discussed above, emotion information is contained in voiced segments [19]. As illustrated in Figure 5, the voice segments have high amplitudes while the un-voice ones have low ones. These features are exploited to select the voice segments. First, to remove noisy spikes, the signal is smoothened using Hamming LPF. Then, the threshold *th* is determined as a function of the mean and standard deviation of the signal. The idea is similar to determining a threshold for removing the silent area, but here it is used to collect the voiced segments. The detailed codes for the Voice Segmentation process are described in Algorithms 3. Since some features require a minimal amount of samples, codes are added to check the length of the segments.

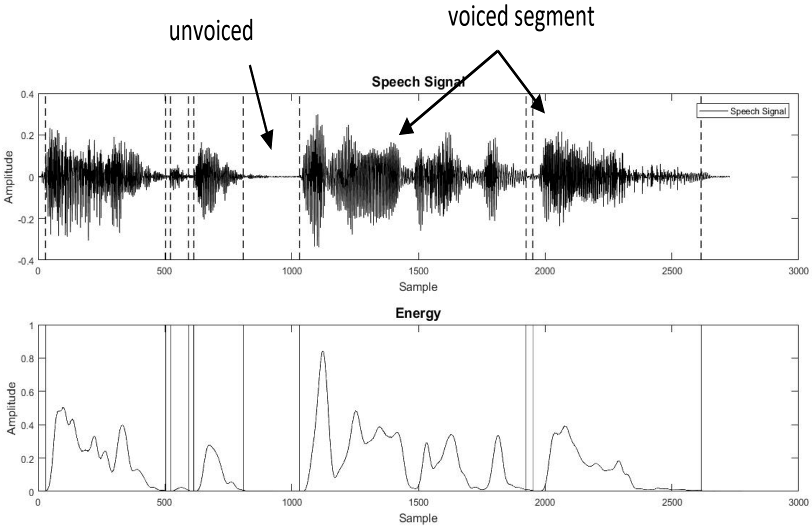


Figure 5. Example illustration of voice segmentation.

Algorithm 3 Voice Segmentation

```
Input: decomposed DWT signal s and its length L
Output : voice signal voice_s(seg_no, sample_no)
Let h is Hamming LPF filter
Let p is allowed length % to be used for discarding short segments
calculate temp equals s^2
energy is temp convolved with h
calculate  meanVal equals mean of  energy
calculate  StDev equals standar deviation of energy
Let th equals β*meanVal /StDev
 for i is 0 to L
    if energy > th then
        set voiced(i) = 1
    else
        set voiced(i) = 0
    endif
endfor
seg_no = 0, sample_no = 0
for i is 0 to L
    if voiced(i) = 1 & voiced(i+1) =1 do
        voice_s(seg_no, sample_no) = s(i)
        sample_no++
    elseif voiced(i) = 0 & voiced(i+1) =1
        seg_no++
        sample_no = 0
    endif
```

```
endfor
% discarding short segments
for $i$ is 0 to $seg\_no$
    if $sample\_no$ on $voice\_s(i, sample\_no) < p$
        remove $voice\_s(i, sample\_no)$
    endif
endfor
reorder $seg\_no$ on $voice\_s(seg\_no, sample\_no)$
```

D. Determining the Number of Voice Segments

Note that the number of voice segments (*seg_no* on Algorithm 3) and the number of samples on each segment (*sample_no*) on each voice signal (*voice_s*) varies because the duration of the individual uttering a sentence varies widely. Meanwhile, the classifier requires a fixed size of input vectors. Therefore, a fixed number of segments is set to accommodate this requirement by finding a maximum number of segments (*seg_max*) among all voice signals in the dataset, expressed by Eq. (2).

$$seg\_max = \max_{0 \le i \le N} (voice_{s_i}(seg_{no})) \qquad (2)$$

where N is the number of voice signals in a dataset. If the original signal has a number of segments less than *seg_max*, the remaining segments on that signal are added cyclically with the original ones. In this work, the necessary segments are added using modulo operation as expressed in Eq. (3)

$$\begin{gathered} n\_voice\_s_i(k, sample\_no) = voice\_s_i(mod(k, seg\_no), sample\_no) \qquad (3) \\ k = 0, 1, \dots, seg\_max \end{gathered}$$

where $n\_voice\_s_i(seg\_max, sample\_no)$, is the voice signal at index i with the same number of segments *seg_max.*

E. Extracting Features

Features extraction plays a crucial role in the overall performance of the speech recognition system [29]. This study applies four existing features (ZCR, Energy, Fourier Coefficients, and Cepstrum) and one new feature (Peaks). The following sections discusses the detailed descriptions of the most common features in the field of speech emotion recognition [4], [29], [30], [31], [32], [33].

1. Zero-Crossing Rate (ZCR)

ZCR of a signal is defined as the rate at which the signal changes from positive to negative or vice versa [4], [29], [34]. It identifies the small changes in the amplitude of a signal to find whether human speech is present in the speech sample or not. Eq. (4) provides the formula of zero-crossing rate:

$$Z = \frac{1}{2} \sum_{k=1}^{M} |\text{sign}(s[k]) - \text{sign}(s[k-1])| \qquad (4)$$

where s_k is the signal's sample at position k, M is the length of the signal, and

$$\text{sign}(s[k]) = \begin{cases} 1 & s[k] > 0 \\ -1 & s[k] < 0 \end{cases}$$

2. Energy (E)

The energy signal is calculated using Eq. (5).

$$E = \frac{1}{N}\sum_{i=1}^{N} TR_i \qquad (5)$$

where *i*: 1…*N*, *TR_i*: trapezoid area between 2 consecutive samples.

$$TR_i = \frac{1}{2}(|s_i| + |s_{i+1}|), i = 1, \cdots N$$

where s_n: $sample_n$

3. Fourier Coefficients (FC)

The Fourier Coefficient (*FC*) retrieves the global information of the frequency content of a signal by converting a function of time into its frequencies using the Discrete Fourier Transform (DFT) [31]. In this study, the Fourier Coefficient is taken from each segment. Eq. (6) is the formula for DFT.

$$|H(k)| = \sum_{m=0}^{L-1} \left| s[m]e^{-\frac{j2\pi km}{L}} \right|, k = 0 \cdots L-1 \qquad (6)$$

where $s[m]$: samples in the segment, $m = 1 \cdots L\text{-}1$ and $|H[k]|$: frequency magnitudes, k=0⋯L-1. L: the number of samples in the segment (varied depending on the length of the segment). To determine Fourier Coefficient (*FC*) from each segment, max from $|H[k]|$ is selected as described in Eq. (7).

$$FC(n) = \max_{0 \le k \le L-1} [|H[k]|] \qquad (7)$$

where n = segment number 1 …. N

4. Cepstrum (C)

The concept of cepstrum has been used in many applications of speaker fundamental frequency and speech analysis [19], [27], [35]. Cepstrum is a tool for investigating periodic structures in frequency spectra. It is achieved as a result of computing the Inverse Fourier transform (IFT) of the logarithm of the estimated signal spectrum as is shown in Equation (8).

$$C = |\mathcal{F}^{-1}\{\log(|\mathcal{F}\{s(n)\}|^2)\}|^2 \qquad (8)$$

where $\mathcal{F}\{.\}$ is the Fourier transform, $\mathcal{F}^{-1}\{.\}$ is its inverse, and $s(n)$ is a speech signal.

5. Number of Peak (P) and the Average of Peak (AP)

A speech signal is characterized by a sequence of peaks that occur periodically at the fundamental frequency of the speech signal. Thus, the maximum peak amplitude during an analysis interval can serve as a simple indication of the amplitude of the signal and as an aid in distinguishing voiced segments from un-voiced ones [35].

This study extracts two peak features: the number of peaks (*P*) is collected from the whole signal, and the average peak (*AP*) is calculated from each segment. Algorithm 4 shows the process of calculating the number of peaks (*P*).

Algorithm 4. Calculate the number of peaks (*P*)
Input: voiced-signal *s,* and its length *L*
Output : *peak*[*i*], number of peaks *P*

```
count = 0
for i = 0 to L
    if $(s[i-1] < s[i])$ && $(s[i+1] < s[i])$ then
        peak[i] = s[i]
        count++
    else
        peak[i] = 0
    end if
end for
P = count
```

Meanwhile, the average peak (AP_n) is taken from the voice segments. So, the available *peak*[*i*], *i* = 0 to *L* extracted by Algorithm 4 is averaged within each segment using beginning-end positions of each segment. The formula is shown in Eq. (9).

$$AP_n = \frac{1}{e_n - b_n} \sum_{j=b_n}^{e_n} peak[j] \qquad (9)$$

where *n*: segment index number, b_n, and e_n are the beginning and end positions of segment *n*, respectively.

As a resumption, the features are extracted from 2 areas: from each voice segment (energy (*E)*, Fourier coefficients (*FC*), and the average of peak (*AP*)), and from the whole signal (zero-crossing (*Z*), cepstrum (*C*), and a number of peaks (*P*)). All the above features build (3**seg_max* + 3)-D vector, which is fed to the classifier, with *seg_max* being the maximum number of voice segments (See Eq. (2)).

F. Classifying Emotion

Since this study focuses on DWT decomposition and voice-part segmentation, it is expected that a simple classifier can be used to prove the scheme's effectiveness. This work adopts the Multilayer Perceptron Neural Network classifier [36] [37], which is composed of three layers: input layer *X* consists of 3**seg_max* + 3 number of features $X \in \Re^{(3*seg\_max+3)}$, the hidden layer consists of three layers with eight (8) neurons $H \in \Re^{3^8}$, and the output layer consists of eight (8) neurons $O \in \Re^8$. Eight (8) neurons on hidden and output layers are selected since the method is validated using eight (8) classes of emotion. Rectifier Linear Unit(*ReLU*) is chosen as the activation function for the hidden layer and *softmax*() at the output layer. Figure 6 illustrates the architecture of our classifier.

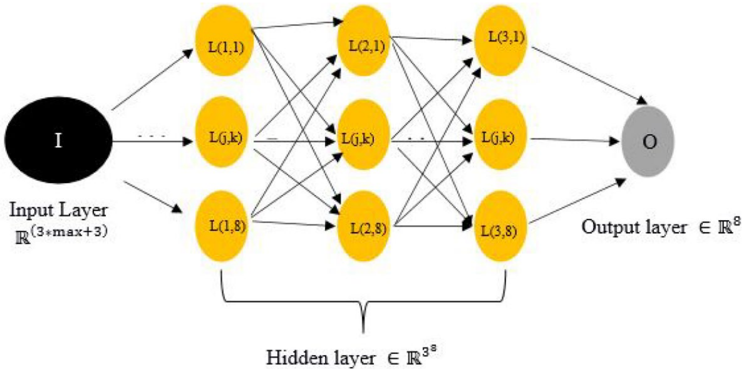


Figure 6 MLP Classifier.

3. Experiment and Analysis

The proposed method is evaluated on RAVDESS dataset in English language [10] which is also carried out by [8], [11], [13], [15], [16], [21]. The dataset consists of 24 speakers (12 males, 12 females) speaking two sentences with eight (8) emotions, i.e., neutral (N), calm (C), happy (H), sad (S), angry (A), fear (F), disgust (D), and surprised (Su). This study only uses speakers' utterances with normal intensity, and the total data are 768 speech signals.

The code for feature extraction is written under MATLAB 2017a. Meanwhile, emotion classification is written in Python programming language with the *sklearn* library. The experiment is conducted on AMD A8-6410 APU, AMD Radeon, 4.00GB RAM, 64-bit workstation.

As a preprocessing step, all silent parts at the beginning and end of the speech signals are removed using the procedures described in *Section 2.A*, followed by DWT decomposition described in *Section 2.B*. At this step, speech signals are at DWT decomposition levels 4, 5, and 6. Then, voice-part segmentations, as described in *Section 2.C* are conducted at all levels of speech signals. Next, from the procedure described in *Section 2.D*, it is found that the maximum number of voice segments (*seg_max*) in DWT decomposition levels 4, 5, and 6 are 18, 12, and 8, respectively. Therefore, based on this finding, the number of features in each level is shown in Table 1.

Table 1. Number of features in each DWT decomposition level

Level of DWT decomposition	No. of features
Level 4	3*18 + 3 = 57
Level 5	3*12 + 3 = 39
Level 6	3*8 + 3 = 27

To validate the model, 8-fold cross-validation is applied. The number 8 is chosen since the dataset consists of 768 speech signals and is distributed evenly into eight (8) classes of emotions. The dataset is divided into eight (8) equal folds, and each fold consists of 3 speakers. In each run, seven folds are used for training, while one (1) fold is used for validation. The scheme is conducted eight (8) times, and the average values of eight (8) results are considered the performance. The formula for accuracy is shown in Equation (10).

$$Accuracy\ (\%) = \frac{TP}{n_{samples}} \tag{10}$$

The experiment is conducted at each level of DWT decomposition, using the feature vector as shown in Table 1. The result is shown in Table 2, which shows that the system's performance at DWT decomposition level 6 achieves the highest accuracy of 98%. As has been mentioned in *Section 2.B*, decomposition level affects the performance of recognition. At each level of decomposition, the noise content is reduced, and the emotion information is better represented by the features.

Table 2. Classification result in each level of DWT

Level of DWT	No. of input features	Accuracy (%)
Level 4	3*18 + 3	66
Level 5	3*12 + 3	92
Level 6	3*8 + 3	98

Furthermore, Table 3 provides a confusion matrix for emotion recognition at DWT decomposition level 6. It shows that some of the calm (C) emotions are misclassified as happy (H) and those sad (S) as angry (A). The case of calm-happy misclassification does not pose a significant problem since calm and happy emotions are correlated. However, in the case of sadness and anger misclassification can be critical. One possible reason is the confusing nature of speech, as the tune of each speaker is almost similar, so it is difficult to identify their emotions. Some emotions are often mistaken for others' since they sound very similar. Emotions could be expressed very differently from one individual to the others. It is, therefore, hard to distinguish emotions between only a few individuals.

Table 3. Confusion Matrix for emotion recognition (%)

Emotion	**N**	**C**	**H**	**S**	**A**	**F**	**D**	**Su**
N	**100**	0	0	0	0	0	0	0
C	0	**92**	8	0	0	0	0	0
H	0	0	**100**	0	0	0	0	0
S	0	0	0	**92**	8	0	0	0
A	0	0	0	0	**100**	0	0	0
F	0	0	0	0	0	**100**	0	0
D	0	0	0	0	0	0	**100**	0
Su	0	0	0	0	0	0	0	**100**

Table 4 demonstrates a benchmark comparison of this proposed method with different *state-of-the-art* SER systems. On the RAVDESS dataset, our system outperforms those of Gao *et al.* [8], Mustaqeem *et al.* [11][13], Zeng *et al.* [15], and Shegokar and Sircar [21]. We infer that their approaches have some weaknesses. For example, [8], [11], [13], [15] used fixed-length audio segments and extracted features from them. This approach has been proven ineffective by the work of [19]. In [21], they also adopted wavelet decomposition, but they extracted their wavelet-derived features from all decomposition levels, which we consider not effective. The highest performance was reported by Biqiao *et al.* [16]. However, their experiment was conducted only on six classes of emotion, while the dataset provided eight classes. They reported that they had conducted a survey to validate the labels in the dataset and concluded that two (2) classes of emotions and one (1) speaker should be eliminated.

Table 4. Comparison of our proposed work with the previous ones

Paper	**Classifier**	**Experiment Settings**	**Accuracy (%)**
Gao *et al.* [8]	Linear kernel SVM with sequential minimal optimization	10-fold cross validation	RAVDESS (6 classes) : 94.5 (highest) EMODB (7 classes): 94.5 (highest)
Mustaqeem *et al.* [11]	Deep Stride CNN classifier	80% training and 20% testing	RAVDESS (8 classes): 79 IEMOCAP (4 classes): 81.72
Mustaqeem *et al.* [13]	Deep BiLSTM network	80% training and 20% testing	RAVDESS (8 classes): 77.02 IEMOCAP (4 classes): 72.25 EMODB (7 classes): 85.57
Zeng et. al. [15]	Gated Residual Networks with Multi Task Learning	5-fold cross-validation	RAVDESS (8 classes) 64.8
Biqiao *et al.* [16]	Directed Acyclic Graph SVM and Regularized Multi-Task SVM	leave one performer and sentence out	RAVDESS (6 classes): 98.19 UMSSED (4 classes): 95.83

Paper	Classifier	Experiment Settings	Accuracy (%)
		cross-validation (RAVDESS) leave-one-performer-out cross validation (UMSSED)	
Shegokar and Sircar [21]	PCA feature reduction, and Quadratic SVM classifier	5-fold cross-validation	RAVDESS (8 classes) 60.1
The proposed method	*Multilayer Perceptron*	*8 fold cross-validation*	*RAVDESS* (8 classes): *98*

4. Conclusion

This work proposes the development of a speech emotion recognition system. The proposed process consists of DWT decomposition and voice segmentation combined with spectral (Fourier Coefficient and Cepstrum) and prosodic features (ZCR, Peak, Energy). Based on the experiment, this method achieves a 98% classification rate accuracy on the DWT decomposition level 6 for the RAVDESS dataset. For future work, it is recommended to use genuine emotions and speaker datasets under a non-controlled environment to obtain better generality.

5. References

[1] A. Dzedzickis, K. A. and B. V., Human Emotion Recognition: Review of Sensors and Methods, Vilnius, Lithuania, 2020.

[2] T. Thanapattheerakul, J. K. Mao, Amoranto and J. H. Chan, "Emotion in a Century: A Review of Emotion Recognition," in *International Conference on Advances in Information Technology*, Toronto, Canada, 2018.

[3] Meribah, Sandhya, Ravichandran and Balasubramaniam, "Silence Removal from Audio Signal Using Framing and Windowing Method and Analyze Various Parameter," *International Journal of Innovative Research in Computer and Communication Engineering,* pp. 1-8.

[4] A. Kaduru, H. Vallveti and A. Budati, "Feature extraction algorithms to improve the speech emotion recognition rate," *International Journal of Speech Technology, Springer,* pp. 1-11, 2020.

[5] P. Ekman, "Emotions Revealed Recognizing Faces and Feelings to Improve Communication and Emotional Life", New York: Times Books, 2003.

[6] Xianxin. Ke, B. Cao, J. Bai, Q. Yu and D. Yang, "Speech Emotion Recognition Based on PCA and CHMM," in *IEEE 8th Joint International Information Technology and Artificial Intelligence Conference (ITAIC)*, Chongqing, China, 2019.

[7] F. Burkhardt, A. Paeschke, M. Rolfes, W. Sendlmeier and B. Weiss, "A Database of German Emotional Speech," in *Interspeech*, Lisbon, Portugal, 2005.

[8] Y. Gao, B. Li, N. Wang and T. Zhu, "Speech Emotion Recognition Using Local and Global Features," in *Brain Informatics,*, e. Y. Zeng, Ed., Beijing, Springer, 2017, pp. 3-13.

[9] W. Eyben and S. B. Wollmer, "Open-Source Media Interpretation by Large feature-space Extraction," Munchen, Germany, 2016.

[10] S. R. Livingstone and F. A. Russo, "The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English," Plos One, https://doi.org/10.1371/journal.pone.0196391, 2018.

[11] Mustaqeem and S. Kwon, "A CNN-Assisted Enhanced Audio Signal Processing for Speech Emotion Recognition," in *Voice Emotion Recognition and Indexing for Affecting Multimedia Service*, 2019, pp. 1-16.

[12] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee and S. S. Narayanan, "IEMOCAP: Interactive emotional dyadic motion capture database," Kluwer Academic Publishers, Netherlands, 2008.

[13] Mustaqeem, M. Sajjad and S. Kwon, "Clustering-Based Speech Emotion Recognition by Incorporating Learned Features and Deep BiLSTM," *IEEE Access,* vol. 8, pp. 1-15, 2020.

[14] C. Sun, A. Shrivastava, S. Singh and A. Gupta, "Revisiting Unreasonable Effectiveness of Data in Deep Learning Era," in *IEEE International Conference on Computer Vision*, Venice, Italy , 2017.

[15] Y. Zeng, H. Mao, D. Peng and Z. Yi, "Spectrogram based multi-task audio classification," *Multimedia Tools and Applications ,* vol. 78, p. 3705–3722, 2017.

[16] Z. Biqiao, G. Essl and E. M. Provost, "Recognizing Emotion from Singing and Speaking Using Shared Models," in *IEEE-Affective Computing and Intelligent Interaction (ACII)*, Michigan, USA, 2015.

[17] E. M. Provost, "University of Michigan Song and Speech Emotion Dataset," University of Michigan, EECS Department, Computer Science & Engineering, [Online]. Available: https://web.eecs.umich.edu/~emilykmp/umssed.html.

[18] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine,* pp. 32-80, 2001.

[19] M. Mansoorizadeh and N. M. Charkari, "Speech Emotion Recognition: Comparison of Speech Segmentation Approaches," in *18th Annual Conference of ICSA Iranian Identity of Curriculum and Instruction*, Tehran, Iran, 2007.

[20] R. Merry, "Wavelet Theory and Applications, A literature study," Eindhoven University of Technology, Department of Mechanical Engineering, Control Systems Technology Group, Eindhoven, 2005.

[21] P. Shegokar and P. Sircar, "Continuous wavelet transform based speech emotion recognition," in *10th International Conference on Signal Processing and Communication Systems (ICSPCS)*, Surfers Paradise, Australia, 2016.

[22] H. K. Palo and M. N. Mohanty, "Wavelet based feature combination for recognition of emotions," *Ain Shams Engineering Journal,* vol. 9, p. 1799–1806, 2018.

[23] P. Jackson and S. Haq, "Surrey Audio-Visual Expressed Emotion (SAVEE) Database," University of Surrey (UK), [Online]. Available: http://kahlan.eps.surrey.ac.uk/savee/.

[24] E. Por, M. v. Kooten and V. Sarkovic, "Nyquist–Shannon sampling theorem," May 2019. [Online]. Available: https://home.strw.leidenuniv.nl/~por/AOT2019/docs/AOT_2019_Ex13_ Nyquist Theorem. pdf. [Accessed 7 september 2021].

[25] A. F. Al-Ajlouni, M. Abo-Zahhad, S. M. Ahmed and R. J. Schilling, "An ECG signal compressor based on the selection of optimal," *Journal of Medical Engineering & Technology,* vol. 32, no. 6, p. 425–433, 2008.

[26] I. Titze, in *Principles of Voice Production*, Prentice-Hall (currently published by NCVS.org), 1994, p. 188.

[27] S. Sunny, D. P. S and K. P. Jacob, "Performance Analysis of Different Wavelet Families in Recognizing Speech," *International Journal of Engineering Trends and Technology,* vol. 4, no. 4, pp. 1-6, 2013.

[28] S. A. Firoz and A. P. Babu, "Wavelet Packets for Speech Emotion Recognition," in *IEEE-International Conference on Advances in Electrical, Electronics, Information, Communication, and Bio-Informatics (AEEICB17)*, Chennai, India, 2017.

[29] J. J. Wolf, "Speech Signal Processing and Feature Extraction," in *Spoken Language Generation and Understanding*, Cambridge, USA, Springer, 1980, pp. 103-128.

[30] R. Banse and K. R. Scherer, " Acoustic profiles in vocal emotion expression," *Journal of Personality and Social Psychology,* vol. 70, no. 3, p. 614–36, 1996.

[31] J. C. V. Correa, "Emotion Recognition from Speech with Acoustic, Non-Linear based Features Extracted in Different Acoustic Conditions", University of Antioquia, Colombia, 2016.

[32] R.Cowie, E. Douglas-Cowie, N.Tsapatsoulis, G.Votsis, K.Kollias, W.Fellenz and J.G.Taylor, "Emotion recognition in Human-Computer Interaction," *IEEE Signal Processing Magazine,* 2005.

[33] M. F. Pradier, "Emotion Recognition from Speech Signals and Perception of Music", Stuttgart, Germany: Institut für System Theorie und Bildschirmtechnik, Universität Stuttgart, Germany, 2011.

[34] Y. Medan, E. Yair and D. Chazan, "Super-resolution pitch determination of speech signals," *IEEE Transactions on Signal Processing,* vol. 39, pp. 40-48, 1991.

[35] R. W. Schafer and L. Rabiner, "Digital Representations of Speech Signals," in *Proceedings of the IEEE*, 1975.

[36] T. Roy, T. Marwala and S. Chakraverty, "Speech Emotion Recognition Using Neural Network and Wavelet Features," in *Recent Trends in Wave mechanics andf Vibration, Lecture Notes in Mechanical Engineering*, Singapore, Springer, 2018, pp. 427-438.

[37] M. Michael T, H. Chandrasekaran and C.-H. Hsieh, "Signal Processing Using Multilayer Perceptron," in *Handbool of Neural Network Signal Processing*, Boca Raton, Florida: CRC Press, 2002.

Hertog Nugroho received B.S. degree in Electrical Engineering from Bandung Institute of Technology, Indonesia, in 1984, M.Sc degree and Ph.D degree in Electrical Engineering from Keio University, Yokohama. Japan, in 1995 and 1999 respectively. He is currently working as Associate Professor in Bandung State of Polytechnic. His research interests include computer vision, pattern recognition, and signal processing.

Raya Nadlira Nurul Fuadiyah received B.S. degree in computer science education from Indonesia education university, Indonesia, in 2015, and M.S. degree in informatics from Telkom University, Indonesia, in 2021. She is currently working as a lecturer in State High School 3, Bandung. Her research interest include pattern recognition, speech signal, and signal processing.