Emotion Sound Classification with Support Vector Machine Algorithm

Chabib Arifin*[1], Hartanto Junaedi[2]
[1,2]Institut Sains Terapan dan Teknologi Surabaya/ Magister of Information Technology
rifinsmpn1ta@gmail.com[*1], hartarto.j@gmail.com[2]

Abstract

Speech is one of the biometric characteristics of human being, as well as fingerprint, DNA, eye retinas. Therefore, no two human beings have exactly identical voice. Human emotion is a matter that can only be predicted through the face of a person or from facial expression alteration, but it turns out human emotions can also be detected through the spoken voice. An individual's emotions like happiness, anger, neutral, sadness and surprise can be detected through speech signals. The development of voice recognition system is still running at this moment. This research analyzes emotions through speech signals. Some related research aims to recognize identity and gender as well as emotions based on conversation. In this research, the writer conducts research on the emotional speech classification on two classes started from happiness, anger, neutral, sadness and surprise. The algorithm used in this research is SVM (Support Vector Machine) with Mel-Frequency Cepstral Coefficient (MFCC) algorithm for extraction. This algorithm contains filter process which is adapted to human's hearing. The result of the implementation process of both algorithms gives the accuracy level of happiness=68.54%, anger=75.24%, neutral=78.50%, sadness=74.22% and surprise=68.23%.

Keywords: *Speech Emotion Classification, Pitch, MFCC, SVM.*

1. Introduction

Voice recognition technology is a biometric technology which does not require great expense and specialized equipment. Sound is one of the unique parts of the human body and can be distinguished easily. Voice recognition is a voice identification process based on the words spoken by the person who captured the sound input device to be recognized and then translated into data which is understood by the computer. When humans emit sound, the sound conveys information by the words spoken through sound waves.

Biometric technology is a self-recognition technique using body parts or human behavior. This technology has two important functions, identification and verification. Identification system aims to solve one's identity. Meanwhile, verification system aims to accept or reject the claimed identity by someone.

Voice recognition technology (speaker recognition) is a biometric technology which is considered as inexpensive and simple. Basically, every human being has something unique/distinctive possessions/characteristics. Sound is one of those unique parts of human body and can easily be distinguished. Automatic emotion recognition and classification on voice signals can be conducted using different approaches such as from text, voice, facial expressions and gestures [1].

Many researchers used different classifiers for human emotion recognition from speech such as Hidden Markov Model (HMM) [2], Neural Network (NN) [3], Maximum Likelihood Bayes Classifier (MLBC), Gaussian Mixture Model (GMM) [4], Kernel Deterioration and K-Nearest Neighbors approach (KNN), Support Vector Machine (SVM)[5] [6] and Naive Bayes Classifier.

In proposed system, basic features of speech signals like pitch, energy, and MFCC are classified into different emotional classes by using SVM classifier**.**

2. Research Method

2.1 Human Voice Signal

Human voice is a signal generated from vocal cord vibration. Sound is a representation of messages to be conveyed by our brain. Human vocal cords vibrate due to the airflow from the

Arifin, C., & Junaedi, H. (2018). Emotion Sound Classification with Support Vector Machine Algorithm. Kinetik, 3(2). doi:http://dx.doi.org/10.22219/kinetik.v3i2.610
Receive February 22, 2018; Revise February 23, 2018; Accepted February 27, 2018

lungs, and from these, a sound wave will produced. The voice produced will depend on the positions of tongue, teeth and jaw or often called articulators, producing certain vowel sounds.

Voice signals are generated and shaped in vocal track. Vocal tract covers an area started from under valve throat (laryngeal pharynx), between the soft palate valve throat (oral pharynx), on velum to in front of the nasal pharynx and nasal cavity. Figure 1 illustrates the area of vocal tract.

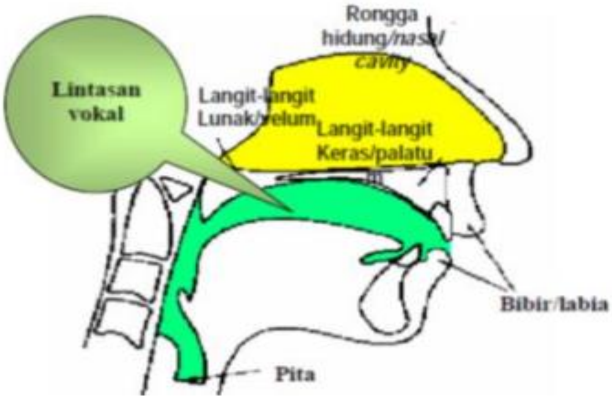
Rongga hidung/nasal cavity
Lintasan vokal
Langit-langit Lunak/velum
Langit-langit Keras/palatum
Bibir/labia
Pita

Figure 1. Vocal Tract [7]

Human voice signal is a relatively long signal which is periodically altered with the speed. Organs involved in speech production process include lung, trachea, larynx, pharynx, vocal cords, mouth, nasal cavity, tongue and lips organ. All can be grouped into three main parts, namely vocal tract, nasal passages, and source generator. The size of the vocal tract varies for each individual, but men own the average length of 17 cm. The size of the vocal tract also varies between 0 (when fully closed) to approximately 20 cm^2. When velum, the organ functioning as the liaison between vocal tract and nasal tract, opens then acoustically nasal tract will join vocal tract to produce a nasal sound.

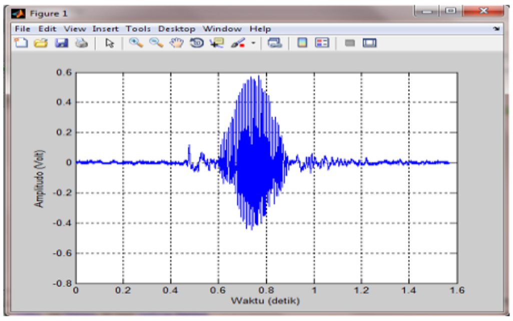

Figure 2. Example of a Human Voice Signal

In Figure 2, the voice signal has a working frequency between 0 to 5 KHz. Components which can be classified from human voice signals are as follows:

1. Region silence, the area when no sound is emitted. Only noise is recorded.
2. Regional unvoiced, the area when vocal cords do not vibrate due to its limp condition.
3. Regional voiced, the area when the first letter of the word is pronounced or when vocal cords have vibrated and already produced sound.

Figure 3 represents the sound signal sampled for 100 ms on each picture. S is an area of silence, U is an unvoiced area and V is a voiced area. However, there are also areas which cannot be categorized included areas experiencing by vocal organ alteration.

Human voice signals which can be heard by human's ear are ranging in area from 20 Hz to 20 kHz. Sound below or above this range will not be heard by human's ear. Human voice has two types, mono and stereo. Mono sound is the sound produced by a single line, so the sound quality is not decent, if it is represented as an image, then mono sound can be represented as gray scale images which only has one layer with the pixel bit. On the other hand, stereo sound is the sound produced by more than one independent audio channel, so the sound is more naturally heard [8].

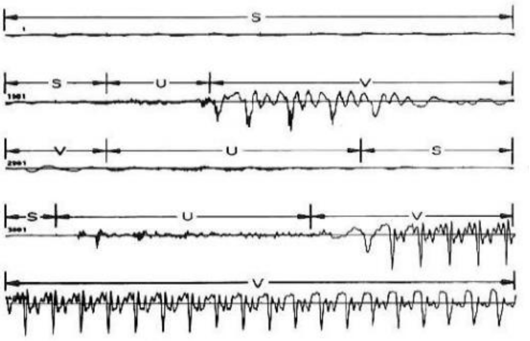

Figure. 3 Snippet of Sound During 500 ms

2.2 Theory of Human Emotion

Emotion is experienced by every human being. An individual’s characteristic may greatly vary from expressing his/her emotion, from changing facial expressions, voice tones and body language. Differences in the existing emotions can be influenced by surrounding environment and people. Emotions may differ according to one’s mood and temperament. Moreover, according to psychological condition, the emotion difference can be felt only momentary. Mood will be altered in a few days. A lifetime temperament of a man will be defined as a human’s characteristics [8].

2.3 Characteristics of Human Sound

2.3.1 Pitch

A wide variety of voices propagated through the air and reflected in all directions can be heard by humans. One of the parameters that can be used to distinguish different types of sound is pitch or fundamental frequency of the sound. Pitch can be defined as the basic tone or sound elements of the smallest human voice. The height discrepancy is low noise associated with the distance between the wave pitch (pitch period) and long-range effect on the frequency. The shorter the distance (meeting), the higher the frequency is. However, it has contrasting effect on the width of the lower frequency range.

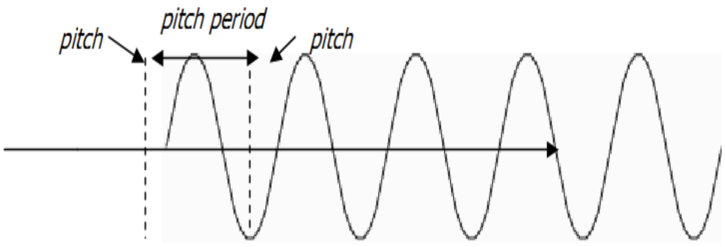

Figure 4. Pitch and Pitch Period

Pitch length area is 10 ms. Human pitch differs depending on age and gender because the vocal cords of women and men have different width which will produce different pitches. Adult males have a lower pitch with the size of the vocal cords is about 17 mm to 25 mm while women have 12.5 mm to 17.5 mm. The intensity of the pitch depends on the tone of voice and the level of human emotion [9].

2.3.2 Energy intensity and duration of pronunciation

In pronouncing a sentence, usually every syllable has a different tone. There are times when the tone should be low or high. A slow or loud spoken by humans is commonly called Energy Intensity. The tone difference is usually desired to give the impression to the sentence pronounced or could be interpreted as our emotional state while speaking these words.

Each individual also has a time discrepancy in saying certain words or phrases. Pauses are required in the so-called word pronunciation with pronunciation duration. There are people who normally take prompt in saying something but sometimes there are mediocre or even require a long duration. It is also influenced by the person's emotional state [8].

2.4 Speech Emotional Features Extraction

Feature extraction is a process for determining a value or a vector which can be used for identifying objects or individuals. In the voice processing, a regular feature value uses cepstral coefficient of a frame. Equation 1 Mel-Frequency Cepstral Coefficient (MFCC) is one of voice signal feature extraction techniques showing with good performance. MFCC is based on the frequency variation limits of human hearing from 20Hz to 20,000 Hz. In other words, MFCC is one type of feature extraction based on the variation of critical bandwidth to the frequency of the human ear. It is a filter which works linearly at low frequencies and works logarithmically at high frequencies to capture the characteristics of phonetic important speech signal. The spectral form of speech signal is used for analysis in spectral analysis [9].

$$f\text{mel} = 2595 \log\left(1 + \frac{f}{700}\right) \tag{1}$$

MFCC is based on the perception of human hearing in which human hearing cannot hear frequencies over 1 KHz. In other words, MFCC shows human ear hearing limit variations with frequency. The entire process of MFCC is pre-emphasizing, framing, windowing, performing DFT, filtering bank, calculating DCT, and delta energy. The block diagram in extracting MFCC process voice signals can be seen in the following Figure 5. Consequently, the following formula can show benefit to compute Mel for a given frequency (Hz) [10]. The block diagram of MFCC extraction processes is illustrated the following Figure 5.

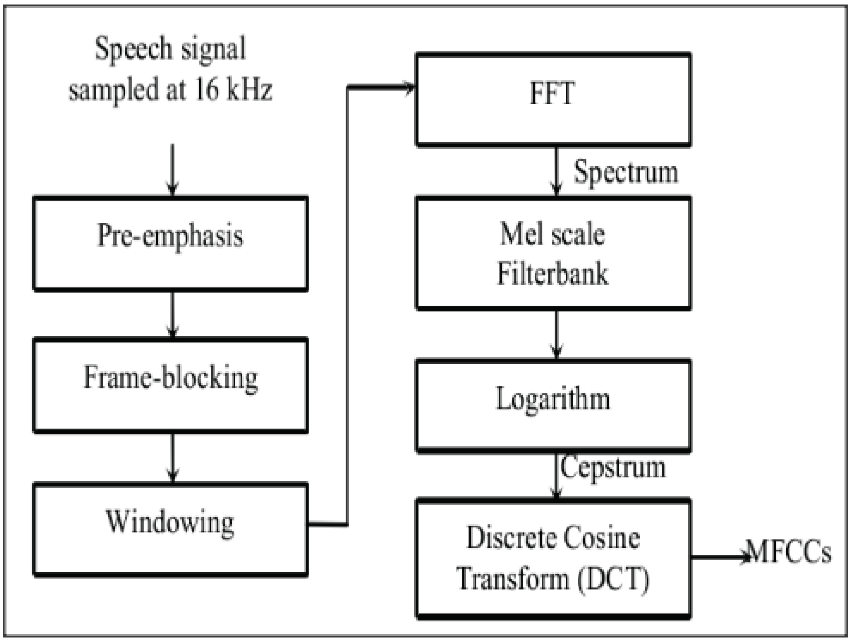

Figure 5. MFCC Extraction Algorithm

Figure 5 shows MFCC extraction algorithm. In the process of MFCC extraction algorithm, how to alter the sound of the linear spectrum signals into sound signals Mel-spectrum non-linear will be explained.

1. Pre-emphasizing

   Pre-emphasizing is a process that is conducted to improve signal quality; thereby, increasing the accuracy at the time of feature extraction. The purpose of pre-emphasis is to spectrally flatten the signal. Z-transform filter is Equation 2.

$$H(\text{z}) = 1 - \mu\text{z} - 1\,, 0{,}94 < \ \mu < 0.97 \tag{2}$$

2. Frame-blocking

   Frame blocking is a process by which signals are divided into frames of N samples of the frames adjacent to the space M. M is smaller than N. This process continues until all of the sound signals can be processed. To calculate the sample point used shown in Equation 3.

   N = sampel rate * waktu per interval

   Where there is a wedge between the sample point:

$$M = \frac{N}{2} \tag{3}$$

3. Windowing

   Windowing is a process to minimize signal discontinuities at the beginning and end of each frame. Hamming window is calculated by Equation 4 for each n samples in each frame.

$$\mathrm{W(n)} = 0{,}54 – 0{,}46 \cos\left(\frac{2\mu n}{n-1}\right), 0 \ \leq \mathrm{n} \leq \mathrm{N} - 1 \tag{4}$$

Then the windowing process is calculated by the following equation:

y (n) = x (n). w (n), 0 ≤ n ≤ N - 1

With y (n) = signal results of windowing samples to - n
x (n) = the value of the sample to - n
w (n) = the value of the window to - n
N = number of samples in frame

4. FFT

   FFT is a fast algorithm for DFT implementation which operates at a discrete-time signal by utilizing the periodical nature of Fourier transformation. FFT is calculated with the following Equation 5.

$$f(n) = \sum_{K=0}^{N-1} y_{k^{e^{-2\pi jkn/N},n=0,1,2,\dots,N-1}} \tag{5}$$

5. Mel-scale filter banks

   After completion of the FFT process. Afterwards, the following step is filtering and goruping the frequency spectrum at each frame, and each band filter will be calculated. A uniformly spaced filter bank at Mel-scale is used for simulating the subjective spectrum. The filter bank filters magnitude spectrum into a number of bands. Low requencies are given more weight than high requency using window overlapping triangles and the number of the contents of each frequency band. This process reflects how the human ear works**.**
6. Logarithm

   The stage of this process illustrates the process of loudness. It can be computed by Mel-frequency cepstral coefficient of the power output from the filter bank using the arithmetic logarithm. This stage maps the logarithm amplitude spectra obtained from Mel-scale as mentioned in the previous process steps.
7. Discrete Cosine Transform

   DCT is the final step of the primary process of MFCC feature extraction. The basic concept of DCT is correlation Mel-spectrum in order to produce a good representation of local spectral properties. Basically, DCT has similar concept with inverse Fourier transform. However, the results of DCT approach Principle Component Analysis (PCA). PCA is the classic static methods which are widely used in data analysis and compression. This has led to often replace the inverse Fourier transform DCT in MFCC feature extraction process. Frequency Cepstral coefficients are real numbers. After the DCT operation, a featured vector with 6 dimensional MFCC is obtained [10].

2.5 Support Vector Machine Classifier

Support Vector Machine (SVM) was firstly appeared in 1992 suggested by Vladimir Vapnik and his colleagues, Bernhard Boser and Isabelle Guyon. SVM is a classification of the types of assisted method (supervised) because when training it requires specific learning targets. SVM can be used for classification which can be applied to handwriting detection, object recognition, voice identification, etc. SVM is an easier and more effective computation technique of machine learning algorithms, under the conditions of limited training data. It is widely used for classification and pattern recognition issues. SVM is a machine learning method which works on the principle of Structural Risk Minimization (SRM) with the aim of finding the best hyperplane which separates two classes in the input space. In contrast to the neural network strategy that seeks class separating hyperplane, SVM trying to find the best hyperplane in the input space. SVM concepts can be explained simply as an attempt to find the best hyperplane serves as a separator of two classes in the input space. In other words, Support Vector Machine is a machine learning algorithm derived from statistical learning theory. The main idea of SVM is to transform the original input into higher-dimensional features using kernel functions and to achieve the optimum level of classification in a featured new space in which there is a clear demarcation between the feature optimal placements of the dividing hyperplane.

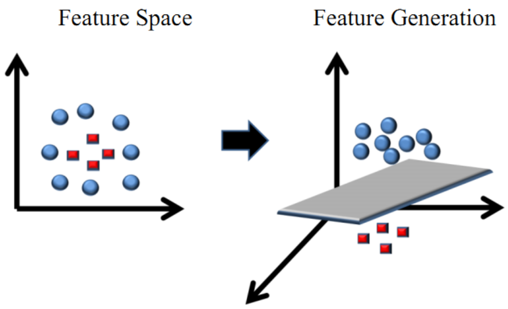

Figure 6. Transformation from Feature Space to Feature Generation

Figure 6 shows a method for classifying data which cannot be separated linearly, with how transforming the data into a feature-dimensional space so that later can be separated linearly through mapping or transformation process [11].

3. Stages of Testing and Analysis of Emotion Classification

To conduct the research process, before recognizing emotions with a good voice, trials of each class with a grade separated between positive and negative will be conducted. Training data will be used as research data in the classification process at the time of testing. Judicial review by voice feature extraction and characteristic classification process. In this study, the voice feature extraction is completed using pitch, energy and MFCC algorithm, followed by conducting the classification process using SVM algorithm. The results of the voice feature extraction would be classified with SVM algorithm to determine the accuracy of the data obtained from the process.

The classification process by using voice emotion will be delivered as follows: The method to classify the data that cannot be separated linearly with how transforming the data into a feature-dimensional space. The process begins with the voice recording conducted by utilizing a microphone. The recording process is determined for approximately 5 seconds. The number of data which will be prepared for this research is as many as 500 units consisting of 10 students aged 15-20 recorded to express all classes (anger, happiness, sadness, surprise and neutral), with each word spoken in class 10 times.

3.1 Pre-processing

Pre-processing stage is the process of inserting the voice data having been saved as * .wav files which were previously conducted using an audacity recording tape as needed. Consecutively, sound signals are filtered into a form which is moresubtle, and information not needed in this process will be removed. Pre-processing stage is divided into three parts, pre-emphasizing, frame blocking, and hamming window. Pre-emphasizing obtains the frequency waveform signal as a more refined sound. Afterwards, after pre-emphasizing, the voice signals are placed into the frame into several parts. After frame blocking, hamming window is conducted to reduce the effects of discontinuity of the pieces or parts of the speech signal.

1. Pre-emphasizing
   Pre-emphasizing is performed to eliminate irrelevant information and noises by using low pass filter calculation. Pre-emphasizing refers to the process of maximizing the signal quality by minimizing the effects such as distorted noise during recording and transmitting data, as well as refining the spectral shape frequency.
2. Frame Blocking
   The sound signal generated from pre-emphasizing results are then placed into the frame into several parts, in which each frame is 30 milli-second long and is separated as far as 20 milliseconds facilitating sound calculation and analysis.
3. Hamming Window
   Windowing is required to reduce the effects of discontinuities of the signal chunks. A windowing method used for processing the speech signal is hamming window by which the sound signal will minimize signal discontinuities at the beginning and end of each frame.
4. FFT
   Fast Fourier Transform (FFT) is a fast algorithm for DFT implementation which operates at a discrete-time signal by utilizing the periodical nature of Fourier transformation. The algorithm is used to evaluate the spectrum of the sound signal by converting each frame into the frequency domain.
5. Mel-scale filter banks
   Filter bank is a technique which uses a convolution representation filter. Convolution can be conducted by multiplying the signal with a coefficient filter bank spectrum. Filter bank can be described with a triangular filter overlap with a frequency determined by the center frequency of the two adjacent filters.
6. Logarithm
   At this stage, this process illustrates the process of loudness. Mel frequency cepstral coefficient of the power output from the filter bank can be computed using arithmetical logarithm. This stage maps the logarithm amplitude spectra obtained from Mel-scale as mentioned in the previously process steps.
7. Discrete Cosine Transform
   DCT is the final step of the primary process MFCC feature extraction. The basic concept of DCT is correlation Mel spectrum so as to produce a good representation of local spectral properties

3.2 Feature Extraction

Feature extraction or feature extraction is an important step in the voice recognition system used in this research to choose the emotion significant features which bring great emotional information about the voice signal. The process finds the voice feature values, wherein the voice feature is gained from pitch and formant. The method used to obtain the value of pitch is autocorrelation, while to get the value of the formant uses a linear prediction coding.

3.2.1 Pitch

Pitch is the fundamental frequency (F0) of the sound signal as the result of acoustic velocity in vocal cord vibration. The greater the vibration of the vocal cords, the higher the pitch value. Pitch period ranges from 10 to 20 milliseconds. Every human being has its own pitch range, depending on the base of the throat of an individual. A typical pitch range (habitual pitch) is recorded by most men from 50Hz - 250Hz, while women have a pitch (habitual pitch) higher than men, ranging from 120 - 500Hz. The fundamental frequency changes constantly and gives someone linguistic information such as distinguishing between intonation and emotion.

3.2.2 Energy Intensity and Pronunciation Duration

In pronouncing a sentence, usually every syllable has a different tone. There are times when the tone should be low or high. A slow or loud speech by humans is commonly called energy intensity. The tone discrepancies are usually employed to give the impression to the pronunciation of the sentence, or could be interpreted as our emotional state when speaking these words.

Each man also has different duration in saying certain words or phrases. Pauses are required in the so-called word pronunciation with pronunciation duration.

3.2.3 Classification

Classification is the process of voice feature data classification wherein the voice feature in this case is the pitch and energy classified by the classification method of support vector machine to obtain sound information generated from both voice features.

In first step, all the necessary features having been previously explained are extracted, and their values are calculated. In obtaining decent voice, the test of training process is conducted prior to recognize emptions. The training will be separated each class.

The training data will be used as research material for the system which can perform the classification process at the time of testing. The test is conducted by voice feature extraction process and the process of characteristic classification. Sound feature extraction is completed using pitch, MFCC and energy as well as algorithms for classification process using SVM algorithm. The results of the voice feature extraction will be classified with SVM algorithm, obtaining the data accuracy being tested.

Testing of the training data uses training data as many as 500. The tests carried out by testing the sound which serves as training data. Testing emotions are performed through voice recognition accuracy testing the accuracy for each emotion. Training data is the result obtained from the feature extraction pitch, energy and MFCC algorithm which will be used for the process of system learning. The classification results of emotion using SVM algorithm are presented in the following Table 1.

Table 1. Classification Results Using Support Vector Machine

Emotion State	Emotions Recognized (%)				
	Happiness	Anger	Neutral	Sadness	Surprise
Happiness	68.54	0	0	16.10	20.45
Anger	15.32	75.24	0	0	16.32
Neutral	0	0	78.50	25.00	0
Sadness	0	0	29.45	74.22	0
Surprise	14.31	19.23	0	0	68.23

The table shows the results of algorithm classification using Support Vector Machine (SVM). Happiness emotion is recorded to have correct test data of 68.54% while the error rate test data is classified as surprise and sadness by 20.45% and 16.10% respectively. The research tests anger emotion which presents correct data of 75.24%, while the error rate test data classified as happiness and surprise by 15.32% and 16.32% respectively. Moreover, in testing neutral emotion, the recorded level of correct data is 78.50% while its error rate test data classified as sadness by 25.00%. In the measurement of sadness emotion, this emotion presents correct data level of 74.22% with its error rate test data classified as neutral accounted by 29.45%.

4. Conclusion

In this research, it can be concluded that the feature classification algorithm SVM (Support Vector Machine) can be applied to the sound classification of emotions with the help of an MFCC (Mel-Frequency Cepstral Coefficient) algorithm for feature extraction.

By using the combined features, the system performance can be improved. The system efficiency depends on an emotional speech sample database used in the system. Therefore, it is necessary to create an emotional speech database accurately and validly.

References

[1] Ritu, D. Shah, Dr. Anil, and C. Suthar, *"Speech Emotion Recognition Based on SVM Using MATLAB,"* International Journal of Innovative Research in Computer and Communication Engineering, Vol. 4, Issue 3, March 2016.

[2] F. Liqin, M. Xia, and C. Lijiang, *"Speaker Independent Emotion Recognition Based on SVM/HMMs Fusion System,"* IEEE International Conference on Audio, Language and Image Processing (ICALIP), pages 61-65, 7-9 July 2008.

[3] R. P. Gadhe, R. R. Deshmukh, and V. B. Waghmare, *"KNN Based Emotion Recognition System for Isolated Marathi Speech,"* Department of Computer Science and IT, Dr. Babasaheb Ambedkar Marathwada University, Aurangabad-431004 (MS) India, Vol. 4 No.04 Jul 2015,

[4] N. Thapliyal and G. Amoli, *“Speech Based Emotion Recognition with Gaussian Mixture Model,” I*nternational Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 5, July 2012.

[5] H. Gang, L. Jiandong, and L. Donghua, *“Study of Modulation Recognition Based on HOCs and SVM,”* In Proceedings of the 59th Vehicular Technology Conference, VTC 2004-Spring. (Vol. 2, pp. 898–902), 17–19 May 2004.

[6] P. Shen, Z. Changjun, and X. Chen*, “Automatic Speech Emotion Recognition Using Support Vector Machine,”* IEEE International Conference on Electronic and Mechanical Engineering and Information Technology (EMEIT), Volume 2, Page(s): 621 - 625, 12-14, Augustus 2011.

[7] Sutikyo, and P. Hadi, *“Sound Processing Based on Age Using K-Means Method*, Surabaya: Surabaya State Polytechnic of Electronics,*”* Sepuluh November Institute of Technology.

[8] R. Magdlena, and L. Novamizanti, *“Simulation and Analysis of Human Emotion Detection from Speech Sound Based on Discrete Wavelet Transform and Linear Predictive Coding*,*”* Faculty of Telecommunication, Telkom University.

[9] Bhaskoro, S. Bagas, Ariani, Irna, and A. A. Almsyah, *“Transformation of Human Pitch Sound Using PSOLA Method,”* ELKOMIKA Journal, Bandung State Institute of Technology, No. 2, Vol. 2, Juy - December 2014.

[10] B. Yu, H. Li, and C. Fang, *“Speech Emotion Recognition based on Optimized Support Vector Machine,”* Journal of Software, Vol. 7, No. 12, December 2012.

[11] A. Rinaldi, Hendra, and D. Alamsyah, *“Gender Recognition from Sound Using Support Vector Machine (SVM) Algorithm*,*”* Information Engineering Study Program, STMIK GI MDP Palembang.