Indonesian Journal of Electrical Engineering and Computer Science
Vol. 21, No. 3, March 2021, pp. 1400~1406
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v21.i3.pp1400-1406

A hybrid strategy for emotion classification

Hussah Nasser Aleisa
Department of Computer Sciences, CCIS, Princess Nourah bint Abdulrahman University, Riyadh, KSA

Article Info

Article history:

Received Mar 25, 2020
Revised Aug 30, 2020
Accepted Oct 1, 2020

Keywords:

Audio-videospeechrecognition in car database
Emotion classification
Emotion detection
Emotion recognition
Support vector machine

ABSTRACT

Human emotion recognition is an upcoming research field of human computer interaction based on facial gestures and is being used for real-time analysis in classifying cognitive affective states from a facial video data. Since computers have become an integral part of life, many researchers are using emotion recognition and classification of data based on audio and text. But these approaches offer limited accuracy and relevance in emotion classification. Therefore we have introduced and analyzed a hybrid approach which could outperform the existing strategies that uses an innovative approach supported by selection of audio and video data characteristics for classification. The research uses SVM for classifying the data using audio-visual savee database and the results obtained show maximum classification accuracy with respect to audio data about 91.6 could be improved to 99.2% after the application of hybrid strategy.

Corresponding Author:

Hussah Nasser Aleisa
Department of Computer Sciences
College of Computer and Information Sciences
Princess Nourah Bint Abdulrahman University, Riyadh, KSA
Email: haleisa2019@gmail.com

1. INTRODUCTION

The facial expressions are assumed to change whenever an emotion is experienced, therefore emotion detection could be achieved by detecting the facial expression associated to it. Facial actions can be extracted from each facial expression. The changes of eyes, mouth and nose positioning could be determined by the movements of facial muscles and computer programs implement the users' facial expressions along with head movements by image capturing approach by representing dots in the coordinate system. The changes are then analyzed as happening of a facial action. There are about 46 facial action units (FAU) found in facial action coding system (FACS) according to a research in 1980 by Ekman et al. By emotion a person is able to communicate and express feelings such as interests, wishes, targets, requirements and much more. Physiological responses are needed in many places of this expression and may change the voice of the person. For instance energy consumption may be more for emotional events like anger (which raises vocal cords vibration, modifies shape and rhythm of the breathing requirements in muscles). Therefore Emotion Recognition represents human voice as data which most of the researchers apply for emotion recognition [1]. In recent studies lot of focus was on this kind of data so as to get better results. So based on speech, text and image the researchers developed many hybrid approaches for emotion classification [2]. It is assumed that voice changes are independent of language and speaker, therefore when classifying emotion, researchers consider only the features of acoustic sound instead of other features as the energy, pitch and the speed of emotional speech changes and mostly the variants because the strong acoustic characteristics of the speech cannot be used individually to determine the emotions precisely and efficiently.

To improve the classification precision of emotions the use of extra voice features like the spectral and prosodic features is done and the video features are then applied as complementary factor. This method would help in enhancing the emotion classification precisely to a higher extent. The aim of this research paper is to develop a mixed emotion classifier using both the audio/video characteristics . To analyze we take into consideration seven types of emotion classes: anger, happiness, fear, hatred, puzzle, shock and neutral. The hybrid strategy estimates the effect of video features usage on the classification precision. The emotions classification is done with respect to audio data only followed by the comparison of results with classification with respect to both audio and video features.

The paper organization initially begins with reviewing the existing work in the area of emotion recognition and classification formulated in Section 2. The Section 3 provides the details of support required for the experimental purpose such as the data collection and related database. Feature extractions is explained in Section 4. The hybrid method is proposed in Section 5. The experimental results are shown in Section 6. Finally, Section 7 concludes the paper and proposes directions for the future research.

2. LITERATURE SURVEY

The use of audio/video (AV) signals or a combination of such signals, to interpret human emotions is a common method to classify and sense human emotion. This paper aims at proposing a good alternative solution to the exixting slutions based on offering increased classification accuracy. There is probability to encounter adverse emotions that add negativity to the emotions. Some of the words such as fright, over stressing, sad and amazement emotions, according to the authors in [3] give negative sense to emotions. On examining the physiological vital activities in humans such as temperature sensing, ecg, blood and air pressure, pulse oximetry, etc within human organs, as was studies by authors [4], the researchers proposed that emotion recognition can be accomplished based on expressions and actions produced by human beings. They also identified and proposed that emotions variations of an individual are responsible for variations in human voice characteristics and this follows that a person could be investigated for his sentiments and emotions based on these Audio Video characteristics. The technique is to determine the basic frequency by mining a persons emotions from his speech [5]. Depending on whether the person is a male or female since the pitch of sound for a male is thicker than female, therefore male frequency range is usually 80-160 Hz and that for female is 150-250 Hz. Small children upto 12 years of age has 200-400 Hz pitch of basic frequency. Predicting a persons' emotion would also be accomplished by other speech characteristics [6] like quickness in speech, its quality and finally the energy criteria either singly or in combined state, thus a structure is created depending on these speech characteristics referred to as Dialect-dependent Speaker Models. Few researchers have covered this dialect determination based on a completely new structure where speaker is considered totally an exclusive speaker who does not apply any knowledge about these features. This anonymous perspective is under research as mentioned by the authors in [7]. The current research in technology developments in the area of audio/speech could enhance the driver concentration by use of audio signaling while driving. If a driver is equipped with an audio interface then it is possible for driver to avoid distractions but there may be noise interference therefore visual data could be appended to improve the user interface by image/video capturing, recording and disseminating both audio-visual data which could be expensive. An AVICAR [8] database contains research dataset for vehicular audio/visual data but due to timelag between audio data and video data streams it is needed to have synchronization since no specific protocol support exists and it's still a matter of exploring new ways to handle this situation.

Such system would offer safety to the driver and other riding people in life-threatening situations. Some Researchers applied the sentiment or emotion identification [9] and hidden markov model (HMM) [10] for categorizing the emotions using the audio signals to devise the results on four emotion classes analyzing happy or angry mood, sadness and also the neutral. For creating and evaluating new models appropriate data usage is important. Many databases are presented for emotion recognition and few are open source [11], eNTERFACEA'05 EMOTION (audio-visual database containing emotional contexts). In [8], a speech corpus database containing multi-channel audio-visual records is provided (AVICAR), given by the researchers in university of Illinois (2003-2004). In [12, 13] there is IEMOCAP1 database for multimodal information capture maintained by SAIL2 lab (University of Southern California).

SAVEE audio-visual database contains data of English speakers who are native males of age 27-31 years and are represented with the labels in [14] KL, DC, DJ and JK. Emotion is designated in separate seven (7) classes: anger, happiness, fear, surprise, sadness, hatred and neutral. The text material in the study consisted of about 120 utterances for every speaker in the 15 TIMIT sentences per emotion utilized. This accounted for an overall size of 480 utterances and the recording of data was made dynamically. An overall 60 painted markers over the actors' frontal face were utilized on facial markers. The distribution of database

in Figure 1 [15] depicts the emotion classes in which columns denote number of video files containing emotional data. It is clearly visible that all emotions are equivalent apart from neutral state.

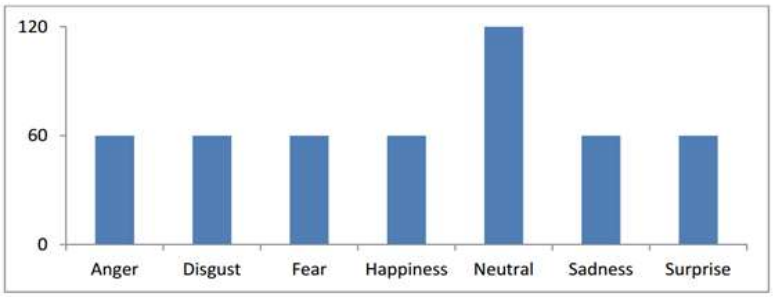

Figure 1. SAVEE Database emotion classes distribution

3. RESEARCH METHODS

3.1. Feature extraction

In this section we review the Audio and Video Feature extraction methods. The Audio Feature Extraction methods are summarized in Table 1 then the Video Feature Extraction follows. Detect Features function of image processing algorithms could be used to extract the feature of an image but the dimension of result obtained is high. Faces are marked by small blue signs in the SAVEE database. The marks are used to identify the essential and effective points in determining a facial expression and to reduce the dimensions of the features extracted. Colour tracking algorithms can be utilized to find these points. A sample of a database image is shown in Figure 2.

Table 1. Summary features of audio feature extraction

Energy and related features	Pitch and related features:	Formant, bandwidth for the first four formants	Mel-Frequency Cepstrum coefficients
This factor is important for speech signals. To obtain the statistics of energy in speech, its value per frame has to be extracted. Therefore the resultant statistics of energy like the maximum value, minimum value, average and standard deviation [6] within the whole speech sample are obtained by evaluating the energy.	Pitch is an important feature in speech emotion recognition. The shape of vocal cords and how they vibrate are affected in different emotional states. Since pitch depends on vocal cords tension and pressure under larynx, and it also contains information about emotion. Pitch signal is called glottal wave-form. Maximum, minimum, average and variation range are different in variety emotion.	formants determination is based on vocal cords that is affected differently in different emotional states. For instance, the highest peak spectral peaks in the spectrum of sound is the first formant frequency. In other words, formant is the concentration of energy around certain frequency. Linear predictive coding method is used for formant calculation [16].	This feature is commonly used in speech with a simple calculation. Mel Frequency Cepstrum has an adequate resolution in low frequency region and has enormous resistance towards noise. But accuracy of emotion recognition is not satisfactory [17].

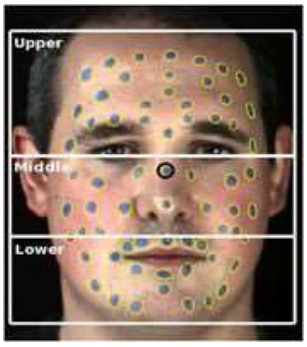

Figure 2. A sample of SAVEE database image and color marker of face

As shown in Figure 2, marker on the edge of nose (encircled in black) is taken as a reference. It is considered as the centre of coordinate, and the remaining coordinates are obtained based on it. With extraction of these features, i.e., by using the same coloured points marked on faces a 130- dimensional set is obtained. This dimensional set includes only three speakers (JK, DC, DJ). It is due to the fact that 65 coloured points are detected on their faces. However, 60 points can be detected and visualized on the fourth speaker face (KL); therefore, extracted features from his related files include 120 dimensions.

To assimilate dimensional features, identify points of difference and their coordinate were removed from other speakers feature files. In Figure 3, the points of difference between the fourth speaker and others (here is JE, the second for example) can be seen. In addition, the second speaker face is marked with yellow circles to compare with the fourth speaker face.

Figure 3. Compare markers of two speakers in SAVEE database

3.2. Emotion classification

Many classification algorithms are being proposed by researchers owing to the wide popularity and applications in Emotion recognition via the audio. Some of these algorithms are found in the following papers [10, 3, 18-20] hidden markov model (HMM), neural networks algorithm (NN), maximum likelihood bayesian classifier (MLC), gaussian mixture model (GMM), kernel regression, k-nearest neighbors algorithm (KNN), and support vector machines (SVM) [21, 22]. SVM structures high-dimension vectors which is actually the hyperplane max distance in the vector space. Although SVM is a simple and efficient algorithm in machine learning, it is extensively being utilized in recognizing issues in classification and pattern recognition. It out performs in terms of classification and the comparison of its performance with other classifiers based on the similar terms when limited training data is easily evident. Therefore SVM is selected as the classifier for emotion classification to analyze the problem under study. The type of SVM (Hard margin) is a non-linear one whereas rhere is another which is an extended version SVM and is referred to as soft margin SVM which contains the definition of a penalty coefficient (C) for data items having class violation. Another SVM, a non-linear is based on Kernel function and separate different classes. Feature space calculation could be expensive based on size and may have unlimited dimensions. Thus, the kernel is used to overcome this problem with RBF kernel function, the use of choice of two parameters is very important in this function. Penalty C in case of conflict and constant σ in Kernel Function [18, 19] are identified, then the classifier can predict emotions comparatively accurately. More than two classes are used in OAA algorithm by generalizing SVM [15, 16].

3.2.1. Emotion classification based on audio features

In this study seven main emotions have been used for identifying emotions: happiness, sadness, anger, fear, disgust, surprise and neutral. Different variations are created in human voice features such as pitch, energy and spectrum in various emotional states. Initially audio features are extracted using one of the feature selection methods. So more effective features can be selected. In this study, z-score method of normalized has been used. Generally, classification uses only audio features with 121 audio features including energy, pitch, formant, coefficient Mel and speed of speech and values associated with them. These features have been selected as they are common in most of works done in this area and the experiments are repeated on multiple choices of audio features. A classification task usually involves separating data into training and testing sets so the classifications were done by 7-fold Cross-Validation and they were examined by different values of sigma (σ) too. In section 6, the results of experiments are investigated.

3.2.2. Emotion classification on audio-visual features (hybrid approach)

In the second phase, classification is done by combination both audio and video features. In other words, part of the work is common with classification based on only audio features. In this phase, extraction of features

must be done first, and the process of dimension reduction must be accomplished. Finally, normalization should take place. The vectors have been created from extracted features which has been used to train SVM in hybrid approach should use audio features vectors and video features vectors simultaneously. A model is created based on SVM classifier and multi-classification done by OAA algorithm. 7-fold cross validation has been used for determining training and testing data sets. The classification accuracy can be improved by appropriate solutions including changes in selected features and checking other ways for feature extraction.

4. RESULTS AND DISCUSSION

The experiments were conducted on MATLAB in which educational algorithms [17] of the University of Rochester were applied to extract audio features. These were applied to improve performance of SVMs, like One-Against-All. Using different kernels had excellent results to solve SVM problems including the application of polynomial kernel [15] in research. They could improve classification accuracy by increasing four percent (4%). Compared to other classification methods, support vector machine method SVM proved to be effective and popular. In this research, the visual features presented in the SAVEE database have been used. There are several tool boxes to easily work for audio and video features extraction such as Open SMILE (and Praat for audio features) and OpenCV for video features that can be used instead of writing algorithm in section feature extraction.

This study uses a limited number of features and has achieved good results compared with other emotion classifiers which are based on audio or audio-visual. For the first stage, classification was done only with audio features in different conditions. Different and remarkable results were obtained. Table 2 shows the result of classification based on just audio features. In this study, we considered σ=9 and three different states of audio features. According to Table 2, if the selected audio features are taken energy and pitch, the classification accuracy is 75.62%. While adding formants to audio features set and retesting, classification accuracy is increased to 82.68%. Next, classification is done with new features set, this time taking into account MFCC and speed of speech in addition energy, pitch and formants features it can be seen that classification accuracy increases based on the result shown and its value of 91%.

Emotion classification was done again with the same audio features, but different values of σ. Values considered for σ were 5, 6.5, 8.5, 9, 10, of which 8.5 achieved the best result. Table 3 shows overall result of models with three values of sigma. Maximum and minimum classification accuracy of seven emotions with only audio features are 91.63% for the model with audio features energy, pitch, formants, Mel-Frequency Cepstrum coefficients, speed and sigma as 8.5 and 75.49 for the model with audio features energy, pitch and sigma as 8.5, respectively. Finally, we wanted to test the impact of adding video features on the accuracy. Thus, we repeated experiments and changed the number of features and sigma with audio and visual features together. The results are summarized in Table 4.

Table 2. Audio classification for seven emotion types

Features Classes sigma (σ)=9	Energy/Pitch	Energy/PitchFormant	Energy/Pitch/ Formant/MFCCs
Sadness	64.97	84.94	93.97
Happiness	73.98	76.87	88.93
Anger	88.53	88.33	93.08
Fear	74.55	85.01	89.10
Disgust	56.37	70.89	89.95
Natural	84.01	87.31	88.99
Surprise	84.99	89.13	93.29
Accuracy	75.35	83.21	90.57

Table 3. Accuracy with only audio features and different amount of sigma (σ)

Audio Features			
(σ) Sigma	Energy/Pitch	Energy/Pitch/ Formant	Energy/Pitch/ Formant/MFCCs
5	81.28	88.77	89.69
6.5	77.49	86.5	91.03
8.5	75.49	83.44	91.63

Table 4. Models accuracy by audio-visual features with different amount of sigma (σ)

Audio-Visual Features			
(σ) Sigma	Energy/Pitch/ VF	Energy/Pitch/ Formant, VF	Energy/Pitch/Formant/MFCCs, VF
5.00	98.85	96.75	90.11
6.50	99.26	97.76	95.03
8.50	99.26	98.81	98.11

The maximum classification accuracy of seven emotions by the hybrid approach is 99.26% achieved from the model with energy, pitch as audio features and video features and sigma as 6.5. Classification accuracy with the same conditions but using only audio features was 77.49. The minimum accuracy was 90.06% obtained from the model with energy, pitch, formants, Mel-Frequency Cepstrum coefficients (MFCC), speed as audio features and video features and sigma as 5. Using the same conditions classification accuracy with only audio features was 89.69%. The comparison of classifications based on only audio features to the hybrid approach (classification on audio-visual features) determines that the hybrid approach increases classification accuracy in all three models. Therefore, the proposed hybrid approach produced more promising results [23, 24] used audio signals for emotion recognition in their work and SAVEE database. Selected features in the project are mainly related to energy, pitch, and statistics and spectral features MFCC as well. They recognized emotions [25] by using linear kernel with binary tree classification strategy "One Against One" (OAO) and "One Against All" (OAs). The best results of that work and this research with the same number of common data and the same audio features are shown in Table 5. By OAO comparing the results in Table 5, a good performance of classification by hybrid approach with the audio-visual features can be seen.

Table 5. Comparing Sinith's project with the proposed hybrid approach

Class	Sinith's work	Hybrid approach
Anger	65	96.55
Happiness	45	98.78
Natural	70	98.78
Sadness	65	97.5
Accuracy	61.25	97.9

The best accuracy in Sinith's work by SAVEE database using linear kernel and binary tree is 61.25%, while the proposed hybrid method based on the same database, exhibits an accuracy equals to 97.90%. In another work given by Chandney [10] and team to recognize emotion which uses Hidden Markov model and SAVEE database referencing four classes: surprise, sadness, fear and disgust. However the work used only one audio feature to recognize emotions, MFCC for which accuracy rate of emotion recognition was 94.17% on the other hand the with same database and same feature the new hybrid approach had raised to 97.82% accuracy. The comparison and results are depicted in Table 6.

Table 6. Chandni's project Vs Hybrid approach

Class	Chandni's work	Hybrid approach
Anger	90	98.24
Happiness	100	99.09
Natural	97	94.78
Sadness	90	99.18
Accuracy	94.17	97.82

5. CONCLUSION

This paper studied various emotion classification techniques and proposed a hybrid technique for classification of human emotions which is the most challenging task in real time situation. We determined our resuls base d on a hybrid criteria that combined audio and video data on a SVM classifier and the improvement on the on the data we used was invincible and ultimately it was seen that the proposed approach outperforms with an accuracy of 99.16 percent, a result more than the research studies currently available in the recent times. This study has given an emotion recognition system which is independent of speaker and language used, also the two audio features considered in the study are prosodic and spectral features unlike the existing researches using different audio features in classification and emotion recognition. The future plan is to investigate emotion recognition with a different perspective of dialect and language impact on emotion recognition. Also we want to experiment the results of recognizing emotions and analyzing audio features in various other languages to check whether it could enhance the accuracy of the classifier. Another improvement would studying the influence of the accent on emotion expression and recognition for audio features. Future research is concerned with the need to create specialized databases for different languages to be considered and then the effective features in any chosen language could be analyzed from the available database.

ACKNOWLEDGEMENTS

This research is funded by was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University through the Fast-track Research Funding Program and I am thankful to the Research Unit for encouraging and giving the women researchers opportunities to do the research in upcoming areas such as Image Processing and make their contributions.

REFERENCES

[1] Ververidis et al , "Emotional speech recognition: Resources, features, and methods," *Speech Communication*, vol. 48, no. 9, pp. 1162-1181, 2006.
[2] Bhaskar, Jasmine, et.al, "Hybrid Approach for Emotion Classification of Audio Conversation Based on Text and Speech Mining," *Procedia Computer Science*, vol. 46, pp. 635-643, 2015.
[3] E. H. Jang, et.al, "Emotion classification based on physiological signals induced by negative emotions: Discriminantion of negative emotions by machine learning," in *Networking, Sensing and Control (ICNSC), 2012* 9th IEEE International Conference on Beijing, 2012.
[4] C. Parlak. and B. Diri, "Emotion recognition from the human voice," in *Signal Processing and Communications Applications Conference (SIU)*, 2013 21st, 2013.
[5] E. Ayadi, et. al, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, no. 3, pp. 572-587, 2011.
[6] Pathak B. V. S., et al.,"Extraction of Pitch and Formants and its Analysis to Identify Three Different Emotional States of a Person" *International Journal of Computer Science*, vol. 9, no. 4, pp. 296-299, 2012.
[7] C. Lijiang, et.al, "Speech emotion recognition: Features and classification models," *Digital Signal Processing*, vol. 22, no. 6, pp. 1154-1160, 2012.
[8] N. Rajitha, et.al., "Recognising audio-visual speech in vehicles using the AVICAR database," in *Proceedings of the 13th Australasian International Conference on Speech Science and Technology Melbourne*, Vic, 2010.
[9] M. S. Sinith, et. al, "Emotion recognition from audio signals using Support Vector Machine," in *IEEE Recent Advances in Intelligent Computational Systems (RAICS) Trivandrum*, 2015.
[10] G. Chandni, et. al, "An automatic emotion recognizer using MFCCs and Hidden Markov Models," in *Ultra Modern Telecommunications and Control Systems and Workshops (ICUMT)*, 2015 7th International Congress on Brno, 2015.
[11] "eNTERFACE'05 EMOTION Database," [Online]. Available: http://www.enterface.net/enterface05/..
[12] C. Busso, et. Al, "IEMOCAP: interactive emotional dyadic motion capture database," *Language Resources and Evaluation*, vol. 42, pp. 335-359, 2008.
[13] A. Metallinou, et al, "Visual emotion recognition using compact facial representations and viseme information," in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010.
[14] "SAVEE Database," [Online]. Available:http://kahlan.eps.surrey.ac.uk/savee/Database.html.
[15] M. Sidorov, et al, "Feature and decision level.audio-visual data fusion in emotion recognition problem," in *Informatics in Control, Automation and Robotics (ICINCO)*, 2015 12th International Conference on Colmar, 2015.
[16] N. Yang, et. al, "Speech-based emotion classification using multiclass SVM with hybrid kernel and thresholding fusion," in *Spoken Language Technology Workshop (SLT)*, 2012 IEEE Miami, FL, 2012.
[17] Y. Pan, et. al, "Speech Emotion Recognition Using Support Vector Machine," *International Journal of Smart Home*, vol. 6, no. 2, pp. 101-108, 2012.
[18] E. Sopov and I. Ivanov, "elf-Configuring Ensemble of Neural Network Classifiers for Emotion Recognition in the Intelligent Human-Machine Interaction," in *Computational Intelligence*, 2015 IEEE Symposium Series on Cape Town, 2015.
[19] S. Agrawal and S. Dongaonkar, "Emotion recognition from speech using Gaussian Mixture Model and vector quantization," in *Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, 2015 4th International Conference on Noida, 2015.
[20] M. R. Mehmood and H. J. Lee, "Emotion classification of EEG brain signal using SVM and KNN," in *Multimedia & Expo Workshops (ICMEW)*, 2015 IEEE International Conference on Turin, Italy, 2015.
[21] N. R. Kanth and S. Saraswathi, "Efficient speech emotion recognition using binary support vector machines & multiclass SVM", *IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Madurai, 2015.
[22] A. Metallinou, et. al, "Context-sensitive learning for enhanced audiovisual emotion classification (Extended abstract)," in *Affective Computing and Intelligent Interaction (ACII)*, 2015 International Conference on Xi'an, 2015.
[23] Y. Chavhan, M. L. Dhore and P. Yesaware, "Article: Speech Emotion Recognition Using Support Vector Machine," *International Journal of Computer Applications*, vol. 1, pp. 6-9, 2010.
[24] M. S. Sinith, et. al, "Emotion recognition from audio signals using Support Vector Machine," in *IEEE Recent Advances in Intelligent Computational Systems (RAICS) Trivandrum*, 2015.
[25] *Fergyanto E. Gunawan, et. al, "*Predicting the Level of Emotion by Means of Indonesian Speech Signal", *TELKOMNIKA (Telecommunication, Computing, Electronics and Control)*, Vol.15, no.2, pp. 665~670 ISSN: 1693-6930, 2017.