International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 34, No. 2, April, 2020, pp. 117-123
Journal homepage: http://iieta.org/journals/ria

MFCC-Based Feature Extraction Model for Long Time Period Emotion Speech Using CNN

Mahmood Alhlffee

Department of DIEC, IIIE, Universidad Nacional Del Sur, Bahía Blanca 8000, Argentina

Corresponding Author Email: mahmood@uns.edu.ar

https://doi.org/10.18280/ria.340201

Received: 20 December 2019
Accepted: 3 February 2020

Keywords:
Mel-Frequency Cepstral Coefficients (MFCC), Fast Fourier Transform (FFT)-Based feature, CNN model and Hybrid HMM/CNN system

ABSTRACT

This paper aims to study the effectiveness of the feature extraction model based on MFCC and Fast Fourier Transform (FFT). Using the CNN model, five basic emotions were extracted from the input speech corpus, and the spectrogram based on long-term speech words was applied to achieve the high-precision performance of the fixed-length learning vector existing in the audio file. Finally, the authors proposed the method of recognizing five emotional states in the FFT-based RAVDSS and SAVEE emotion speech corpus based on FFT. By comparison with the most advanced correlation methods, it's found that the detection accuracy is improved by 70% when using the proposed model to extract audio fragments from audio files and adjust the speech words to spectrograms.

1. INTRODUCTION

Speech or audio signal is one of the most sentimental and natural way for communication among human beings. Emotions make speech more expressive and effective. Humans use different speech tones to express their emotions [1]. Therefore, the emotion categories fall into the different levels Like, Happy, Normal, Sad, Angry, and Disgusting. The speech emotion detection can be an easy task for humans but it is a very difficult task for machine. This kind of difficult has motivates huge number of the researchers to consider speech signal as high effective way to human-to-machine interaction. It means to allow computer to interact appropriate to human intentions and make human computer interaction quite easy [2]. However, the recognizing of emotional conditions in speech signals are very challengeable task due to defining the boundaries and distinguishing between them and measuring them in a meaningful way. To our best knowledge, no one considers a long time period audio signal detection to generate a more appropriate match system.

The ASR is an independent method for communicating human-to-machine based on speech signal as shown in Figure 1. The ASR system aims to enable natural human-computer interface to extraction of the emotional state of the speaker tone language from the input speech signal of the users. Therefore, to implement such system, many core technologies include Hidden Markov models (HMMs), Gaussian mixture models (GMMs), Mel-frequency cepstral coefficients (MFCCs), etc. [3, 4]. The general concept of ASR system is typically having several architecture layers, which are optimized in an independent manner [5]. In the first architecture layer, is the raw signal that transformed into features. A feature that composed of two typically phases the dimensionality reduction and the information selection phases, those phases are based on task-specific phenomena knowledge, that leading to state-of-the-art features. In the second architecture layer, the likelihood of sub-word units such as phonemes are estimated using discriminative or generative models. In a final architecture layer, dynamic programming techniques are used to recognize the word sequence given the lexical and syntactical constraints [4, 5].

The MFCC technique is the one of the most popular spectral based parameters used in recognition system approach. The MFCC has advantage over servals techniques which is less complexity in implementation algorithm for feature extraction. The MFCC feature extraction technique basically includes several filters start from windowing filter, Discrete Fourier Transform filter, log filter for the magnitude, etc. The MFCC extracted all the samples and then statistically analyzed the principal components, at least 2D minimally required in further recognition performance evaluation [6].

The advantage of the recent deep neural network has made a possible system that can be trained as end-to-end manner, a system where every step is learned simultaneously. The CNN, is the one of few successful neural network architecture for such tasks which has proposed for time series data as the convolution-based filtering is performed along with time axis to help solving recognizing objects of the speech recognition problem [7-9].

The aim of this paper, is attempt to solve the issue of the emotion speech classification in speech audio in-order to obtain high accuracy performance classification and minimize the computer resource demands. That is, our goal is to investigate the effectiveness of MFCC filters technique in long period speech raw signal and understand the features that could improve the system accuracy performance, then use of CNN model to analysis and understand the signal speech information. To that end, our method implementation with less complexity structure and easy to work with a different datasets type to compute the mean frequency responses of the filters in the first convolution layer that match to the specific inputs representing. Our studies on RAVDSS and SAVEE emotions speech corpus task is to segment the audio signal that extracted from the audio files and resize the speech-word into spectrogram forms *(8ms-18ms)* speech signal, then present a study to evaluate the mismatched conditions based on CNN system model.

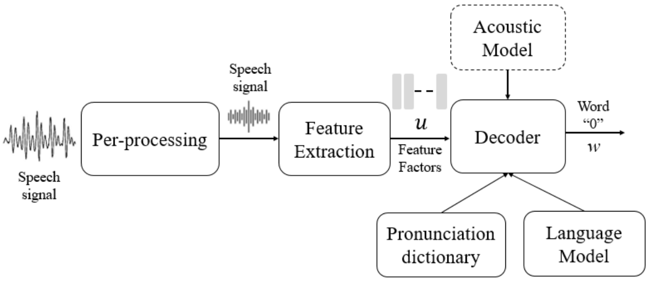

Figure 1. Automatic speech recognition system

The remainder of this paper is organized as follows. Section 2, a brief summary of the related frame-work. Section 3, a brief discerption of speech recognition system architecture. Section 4, a brief discerption of speech coupes. Section 5, the evaluation result of our model. Section 6, the conclusion.

2. LITERATURE RELATED TO THE FRAMEWORK

For last two decades, a huge study has been conducted regarding speech recognize emotions by using speech statistics. Deep learning has been proved to have powerful ability for emotions recognize task. Many publications work have been proposed a different deep learning neural network architecture to overcome a better accuracy performance such as Cao et al. [10] proposed based on SVM ranking method for emotion synthesize information recognition to solve the binary classification issue. However, the SVM ranking method is only for particular emotions, treating data from every utterer as a distinct query then mixed all predictions from rankers to apply multi-class prediction. The reason behind Cao et al work within three level speech emotion recognition method, was to improve the emotion recognition system in speaker-independent. Another work from Nwe et al. [10, 11], proposed a new emotional classification system for utterance signals. They idea was implemented by used of LFPC and discrete HMM to characterize the speech signals and classifier respectively. Their method was based on two stages; First, they classified the emotion into six different emotion categories, Secondly, they trained and tested the new system via used of private dataset. And Hence, they proposed a new technique called Modulation Spectral Seatures (MSFs). Another work from Rong et al. [12], presented a method with a high number of features called ERFTrees (Ensemble Random Forest to Trees) for emotion recognition without referring to any language or linguistic information. However, it is worth mention that the language or linguistic information still an open problem in emotional speech recognition system so-far. In work of Narayanan et al. [13], they proposed a domain-specific emotion recognition, their experiment used of a call center application as database to utilizing speech signals to extract different types of information include acoustic, lexical, and discourse for detecting negative and non-negative emotional states with main focused on few emotional like happy, sad and anger. Albornoz et al. [14, 15], their work was to investigation the new spectral feature emotions and to characterize groups. Their investigation was categorized into three different stages a novel hierarchical classifier, acoustic features and grouped emotions to design a novel hierarchical technique for emotions classification. Moreover, to evaluate this investigation three core classifier technique involved GMM, HMM and MLP with distinct configuration and input features. Lee et al. [16], their work was to identify the emotions via used a hierarchical computational structure method. Their work method was categorized into two stages a following binary classifications layers and maps the input audio signal in one of the corresponding emotion classes. The main concept of this method makes an easy way to solve the classification task and to minimize error propagation. Dai et al. [17] proposed a concept of computational approach for emotion recognition and analysis the specifications of emotion in voiced social media which is mixture of GMM autoregressive for the emotional speech classification problem.

2.1 MFCC method for speech recognition

Speech recognition is the ability of a machine or program to automatically recognizing the spoken words or spoken language and convert or phrases them into a readable-format that the machine can understand. With use of recognition technique nowadays, it makes it possible for this technique to be used in verifying and identity the user's voice for servers like voice mail, information service, access the database services, voice dialing, Telephone banking, etc. [18]. Therefore, many core technologies have involved to implement such system include MFCC, PLCC, PCA, etc. However, the MFCC method is considered to be one of the best among others with less complexity in implementation for the feature extraction algorithm. The MFCC use only sixteen coefficients to corresponding to the Mel scale frequencies of speech cepstrum that extracted from spoken word samples in speech coupes, then all the MFCC extraction samples are statistically analyzed for principal components and converted at least into 2D for performance evaluation purpose. In general, the MFCC architecture has several filters stage representation for the feature extraction process which are elaborated in Figure 2.

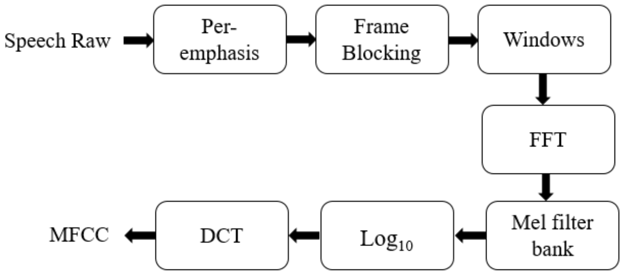

Figure 2. MFCC feature extraction

2.2 Filter model

The MFCC feature extraction technique basically is a compile of several stages of filters which is applying to a speech raw signal.

Pre-emphasis: The purpose of per-emphasis filter is to balance the spectrum of voiced sounds by boost a high signal frequencies range, while leaving the low frequency range in their original state. In most case scenario, the pre-emphasis filter is use to removes some of the effects call distance issue or octave slope range with a true spectrum of the vocal truce from vocal tract parameter which cased via microphone [19]. The pre-emphasis filter is calculating by the given function:

$$H\ (z) = \ 1 - bz^{\ -1} \tag{1}$$

where, *b* presented the value that controls the filter slope with range between *0.4 and 1.0*.

Windowing and Frame blocking: This type of filter is use for a quasi-stationary signal or slowly time-varying. The purpose of windowing filter is to stabilize acoustic characteristics and divide the signal into equal number of short-term frames. The short-term frames are segmented into frames-size of *20 ms* to *30 ms* with optional overlap of *1/3~1/2* of the frame size. This frames-size range is to allow the temporal characteristics of individual speech sounds to be analysis and tracked to help resolve significant temporal characteristics issue and provide good spectral resolution. Hamming or Hanning is the type of windows are generally use in this filter [20].

Mel-Frequency Cepstrum (MFC) or Mel-spectrum: A Mel-frequency cepstrum is a representation a sound with a short-term power spectrum unit that use to measure the human ears perceived frequency. The Mel-frequency cepstrum is based on linear cosine transform of a log power spectrum on a nonlinear Mel scale of frequency. correspond linearly to the physical frequency of the tone, as the human auditory system apparently does not perceive pitch linearly [21]. The Mel scale is calculating by the given function:

$$f_{mel} = 2595\, log_{10}\left(1 + \frac{f}{700}\right) \tag{2}$$

where, *f* denotes the physical frequency in *Hz*, and f_{mel} denotes the perceived frequency. In above function, the linear frequency spacing range is below *1 kHz* and the spacing of logarithmic above *1 kHz*.

For a MFCC technique computation, the filter bank can be implemented in both time domain as well as frequency domain. A triangular shaper filter or Hanning filter among the most commonly used filter and can be calculating by the given function:

$$s\,(m) = \sum_{k=0}^{N-1} [\,|X(k)|^{\,2}\, H_m\,(k)]\,;\; 0 \le m\; \le M\; - 1 \tag{3}$$

where, *M* presented as a total number of triangular Mel weighting filters.

Discrete Cosine Transform (DCT): Is a technique that applied to the transformed Mel frequency coefficients produces a set of cepstral coefficients. For task like speech recognition the DCT is used to generate a *2D* time matrix to recognize each pattern and minimize the size of the speech information. DCT is used in system for speed-up by remove the redundancy from audio information, and hence, the system can be made robust by extraction only those coefficients ignoring or truncating higher order Discrete Cosine Transform component. Therefore, the DCT can be calculating by the given function:

$$c\,(n) = \sum_{m\,=\,1}^{M\,-\,1} Log_{10}\,(s(m))cos\left(\frac{\pi n(m - 0.5)}{M}\right);\; n = \;0\,,1\,,2\,,\ldots,C\; -\; 1 \tag{4}$$

where, *c (n)* presented as cepstral coefficients and *C* presented as a MFCCs number [20].

3. DEEP LEARNING MODEL FOR SPEECH RECOGNITION

The advantage of the recent deep neural network has made a possible for the system to use deep learning techniques in ASR system for feature extraction and language modelling. The algorithms method of the deep learning has the capability of enhance the computers system to understands the raw input speech signal. Gaussian mixture models (GMM), is the one of the most effect technique that use to representing speech signal in the speech recognition system. GMM technique method is based on Hidden Markov models (HMMs), the theory behind that is the speech signal can be considered as a short time stationary signal or piecewise stationary signal.

3.1 Hybrid system

The hybrid system architecture based on ASR is composed of three stages: the MFCC features extraction stage, classification stage and decoding stage as shown in Figure 3. At the MFCC stage, the features are extracted from the raw signal, by use of filtering and transformation techniques method. Then, the extracted signal with some context of four frames are fed as input to the CNN [22-24]. The CNN neural network is usually a feed-forward MLP which categorized into three layers; input layer, hidden layer and output layer which estimates the conditional probabilities for each phoneme class.

The sequence of audio speech files is provided as input signals, then the input signals are divided into window frames with score sign level for each class and each frame by use of MFCC at the first part. At the second part of the system architecture, the CNN network is provided with multi-filters feature learning, followed by a modeling stage.

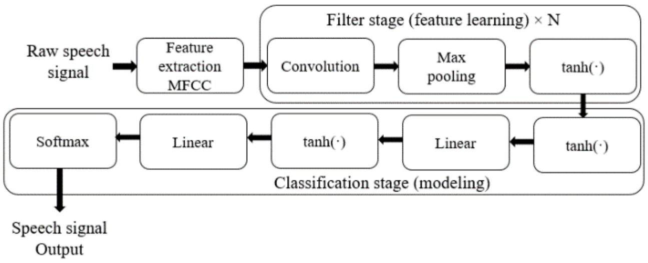

Figure 3. Hybrid system model architecture

4. SYSTEM ARCHITECTURE

There are some methods that applied to the speech signal in this work. A MFCC / Fast Fourier Transform (FFT) based-method filter that takes the raw speech signal as inputted signal as shown in Figure 4. In this work, the FFT method considered to be a standard representation of a speech signal in the frequency domain [24]. However, the disadvantages of such method is that, it cannot be fit for a signals whose frequencies are time varying, in most cases, the signals are assumes to be stationary in nature, in order to allow working in frequency domain, then used of the speech signal frequency spectrum as a waveform substitute. Moreover, working in frequency domain can be helpfully in order to provides more information regarding the speech signal and distinguish between speakers.

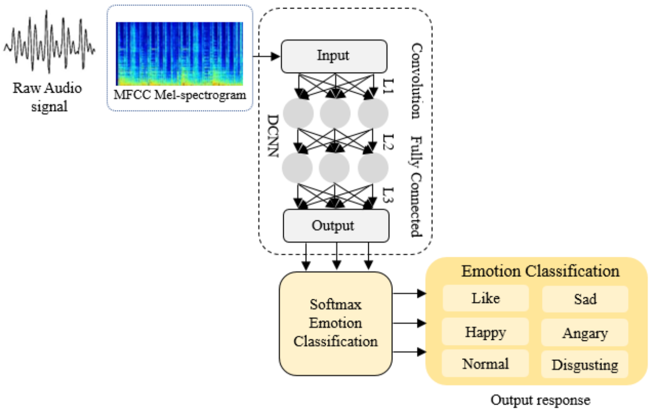

Figure 4. Speech recognition system architecture

FFT Principle: the principle of FFT it is used as conversation method to convert the time signals $x\ (j)$, into the frequency $x\ (k)$. The N here presented as complex sequence numbers of $x_0, \dots, x_{N-1}$ in time signal domain [25, 26]. The FFT can be calculating by the given function of $x\ (j)$:

$$X\ (k) = \sum_{j=0}^{N} x\ (j)\ e^{\frac{-2\pi i}{N}(j-1)(k-1)} \qquad (5)$$

where, $k = 0, \dots, N-1, x(j)$ presents the sample at time index of j and $\sqrt{-1}\,.\,X(k)$ is a vector of N values at frequency index k corresponding to the sine wave magnitude that resulting from the time indexed signal decomposition. The inverse FFT can be calculating by the given function:

$$x(j) = \frac{1}{N}\sum_{j=0}^{N} X\ (k)\ e^{\frac{(-2\pi i)}{N}(j-1)(k-1)} \qquad (6)$$

The advantage FFT method in symmetry and periodicity properties is to reduce computation time [26, 27]. However, this type of method has some performance limitation such as complexity algorithm transform when operates on an imaginary number.

4.1 CNN model

In this work, our neural networks architecture CNN model came after the MFCC stage. The CNN take raw speech which slides into spectrograms via MFCC as shown in Figure 3.

Convolutional Layer: In this layer architecture, our model fed with a sequence of T vectors / frames: $X = x^1, x^2 \dots x^T$. In this layer, we implemented the same linear transformation over each dW frames interspaced (or) successive kW frames window as shown in Figure 5. This layer can be calculating by the following transformation at frame t with a given function:

$$\mathrm{M}\begin{pmatrix} \mathrm{x}^{\mathrm{t-(kw-1)/2}} \\ \vdots \\ \mathrm{x}^{\mathrm{t+(kw-1)/2}} \end{pmatrix}, \qquad (7)$$

where, M presented as parameters matrix of $d_{out} \times d_{in}$.

Max Pooling Layer: In this layer architecture, the model performs a local temporal max operation over a sequence of input, as shown in Figure 6. This layer can be calculating by the following transformation at frame t with a given function:

$$\max_{t-(kW-1)/2 \,\le\, s \,\le\, t+(kW-1)/2} x_s^d \qquad (8)$$

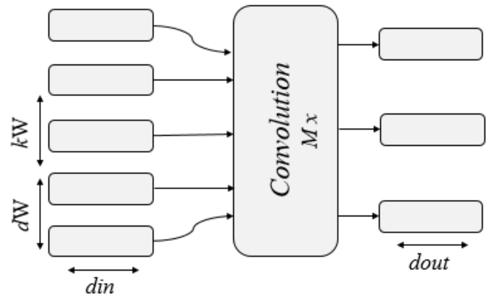

Figure 5. The d_{in} and d_{out} presented the input and the output dimensions' frames, whereas Kw (kW = 3) presented as the width kernel and dW (dW = 2) presented the shift between two linear applications

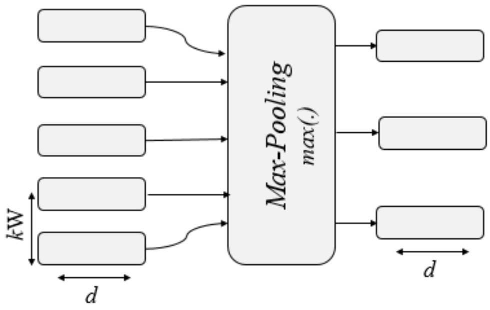

Figure 6. The kW (kW = 2) presented as frames taken number for each max operation and d represents as input / output frames dimension (which are equal)

with x presented as input while d presented as dimension. The purpose of this layer is to increase the robustness of the network.

Softmax Layer: In this layer architecture, the method of Softmax is implemented at the latest layer instead *tanh*, *ReLU*, *sigmoid* or even another neural network activation function. The reason behind the use of Softmax is to converts the output into essential probability distribution and interprets as conditional probabilities [26, 28]. This layer can be calculating by the following class label I with a given function:

$$p\ (i|x) = \frac{e^{f_i\,(x)}}{\sum_j e^{f_j\,(x)}} \qquad (9)$$

4.2 Speech dataset model

This experiment used two different type of the raw speech signal corpus of five emotional states in the RAVDSS and SAVEE. The corpus was in "WAV file format". The coupes contain *400 to 1000* recorded files with a emotions speech in English language with very less noise generating. The goal of the long period emotions speech time is to test our model for a possible improvement design. Moreover, in this work the model can provide supporting when the audio segment was extracted from the audio files of a long time speech period and resize into spectrogram forms. The two different types of coupe are used for training, testing and evaluating our model performance on five emotional states based. However, for the training purpose, our model provided with one sentence words for each speech file.

4.3 Speech database model

There were two type of the database used in this experiment. The first type of database used as a test set while the second database used as train set. During the data setup, our model was augmented with some addition white Gaussian noise of *+15 SNR* to each file of the speech signal either *10 times* or *20 times*. Therefore, we have two sets of data augmentation; set of *10 times* data augmentation *(10x)* and set of *20 times* data augmentation *(20x)*. For test purpose, our model provided with original type of data without noise. All the training data were labels at the end while the test data were encoded as one-hot vectors. The MFCC / FFT filter helped to resample all the audio files at the frequency rate prior of *16 kHz* to any processing. Therefore, all the audio files were then converted into spectrograms form as shown in Figure 7 and Figure 8. A spectrogram is an image that displays with two axes dimensional the horizontal axis (abscissa) represents as time and the vertical (ordinate) presents as frequency. The level of energy or intensity of the spectrogram are spotted or encoded by different level of colors or darkens. The spectrogram has two categorized types; spectrograms with a wide-band and spectrograms with narrow-band. However, the different between them is that, the time resolution property of spectrogram with a wide-band is higher than spectrogram with a narrow-band, while the frequency resolution property of spectrogram with a narrow-band is higher than spectrograms with a wide-band. Therefore, the spectrograms with a wide-band enables to show individual glottal pulses. Meanwhile, the spectrogram with a narrow-band enables to resolve individual harmonics.

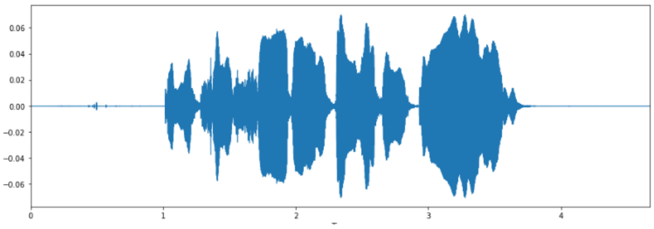

Figure 7. Speech signal form

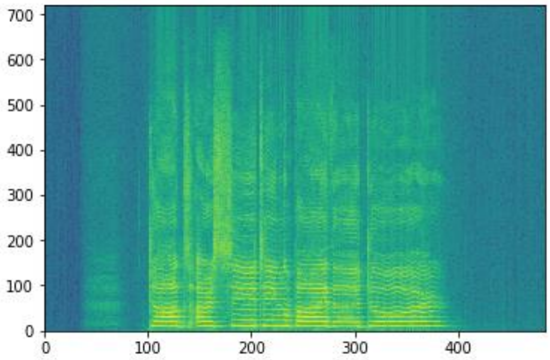

Figure 8. Spectrogram form

The purpose of data augmented, is for training used only. The final stage, is to labels of the training data while the test data were encoded as one-hot vectors. Table 1, summarizes the labels that assigned to each database on their corresponding accuracy on the test data. The epochs number of training data was setup between *400 to 1000*. However, in order to minimize the system resources and time expenses we reduced the training set up to ***300 epochs***. In our model, we ran a single language dependent, with several gender-independent experiments on each database.

Figure 9 and Figure 10 demonstrate the CNN model accuracy over trained and tested database for each iteration of learning. Therefore, our observation shown that, the relationship in the performance of our network during the 300 epochs between training and testing dataset.

Table 1. Emotional predicted

	Actual values	Predicted values
88	female_angry	female_angry
89	male_fearful	male_happy
90	female_fearful	female_happy
91	male_calm	male_calm
92	male_fearful	male_happy
93	female_angry	female_angry
94	male_happy	male_calm
95	male_angry	male_angry
96	male_sad	male_sad
97	female_calm	female_calm
98	male_angry	female_happy
99	male_sad	female_calm

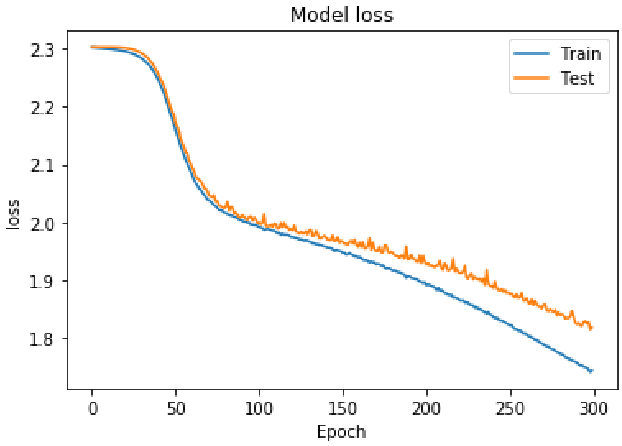

Figure 9. The average loss performance for trained and tested over two different databases

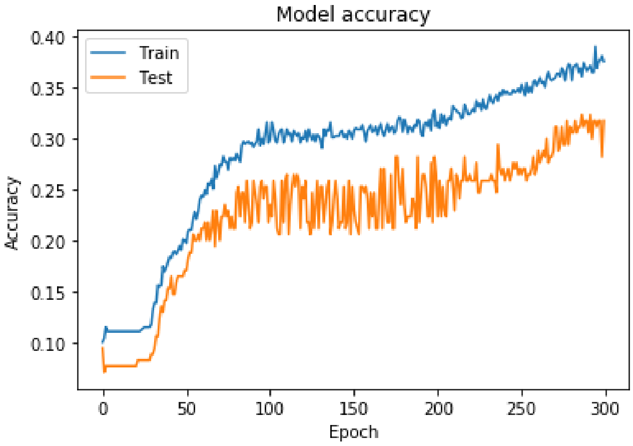

Figure 10. The model accuracy performance for trained and tested over two different databases

5. MODEL EVOLUTION RESULT

Our evolution result is to classify five different emotional categories from the input speech audio and analysis the accuracy performance of the speech emotion recognition system. Moreover, this evaluation is to detect the emotional state from the long speech period of an audio signal and model compare the functionality of MFCC / FFT of identify the five basic emotions with many exiting state-of-the-art systems which are designed to provide better performance when using long speech period time. Based on this comparison, our model can show the average percentage of how much detection information can detect from input raw speech signals and help to provide a better result as compare to existing state-of-the-art model as illustrated in Table 2. The system architecture is a compile of several different components. First, our model used the MFCC to compile the audio segment files into spectrogram forms to obtain high resolution 2D image which is carrying both syntactic and semantic information, then use of that information to feed the CNN layers. The proposed FFT-model achieved a 70% detection accuracy when the audio segment was extracted from the audio files and resize the speech-word into spectrogram forms. Therefore, it is worth mention that, the model can help to improve and understand the use of MFCC / FFT on audio file for resize audio speech-word into a spectrogram form. Our model can show the capability of extract several emotions from a raw speech signal and improve the performance result overall.

Table 2. The proposed method compares to several existing methods in similar relevance

Different feature extraction methods for the speech recognition system	**The accuracy performance (%) over different emotional speech corpus**
The proposed method	The accuracy range (65% - 75%) on labelled datasets detection when the audio segment was extracted from the audio files and resize the speech-word into spectrogram forms. The model helped to improve and understand the use of MFCC / FFT on audio file to resize the speech-words into a spectrogram form. The MFCCs technique help to implies the feature vectors from the frequency spectra of the windows speech frame, transformation of long time speech period to frequency domain. Moreover, the MFCC provide some advanced technique including non-linear technique and transformation technique which used to reduce the dimension decorrelated features.
Principal Component Analysis (PCA) feature extraction [22-25].	The PCA technique is often used for data reduction / compression and minimize the risk of loss the information. PCA is used only to provide information on the true dimensionality of a data set, due to its linear calculation nature, LPC could not extract noisy signal at high amplitude, take a long time to extract the features and used as a one of the common short-term spectral measurements. The are two-independent popular PCA algorithm methods for feature extraction. However, the drawback of these methods is that, their optimization criteria are different from the classification minimum error criterion. These criteria different may lead to inconsistency case between the feature extraction and the classification stages of a pattern recognizer. The final fact about PCA is that, it cannot include a priori information on the speech signal under test.
Linear Predictive Cepstral Coefficient (LPCC) feature extraction [26, 28, 29]	LPCC is a technique that used to capture the vocal tract characteristics a specific emotion information from input speech signal, and cannot include a priori information on the speech signal under test, LPCC is highly susceptible to the quantizer noise and requires the use of proper ordering, LPCC is use to analysis on a high-pitch speech signal gives small source-filter separability in the frequency domain worked at low bit-rate and represented an attempt to mimic the human speech.

6. CONCLUSION

ASR is remaining a challengeable task, due to them interpret signal and recognition pattern. In this work, the investigated on the ASR performance based on CNNs model, which takes raw speech signal and feed to FMCC then input to CNN model. This work, we implemented CNN to classify emotional states using of RAVDSS and SAVEE emotional speech corpus. All raw speech signals were converted into spectrograms wide-band and fed to the CNNs as the inputs. However, most of the related work in the emotion recognition system were use of narrow-band spectrograms from, which have higher frequency and resolve individual harmonics than the spectrograms with wide-band. The proposed approach of this work, is to spontaneous emotional state considering long period time of the emotion speech only. This study proposed a well-known technique frequency based-on feature extraction called MFCC, which is commonly used to improve the speech signal classification performance. Moreover, MFCC and HMM recognition based methods are varying under different effective environment circumstances, but in the real world application the ASR system still suffer due to some major limitations. In conclusion, the accuracy results shown the effects of MFCC feature extraction methods when it is implemented as first step of the entire system.

Our plan in the feature, is to design and employ speech emotions recognition system considering empathy dimensions for the input speech signal by GAN and CNN model.

REFERENCES

[1] Hess, U., Thibault, P. (2009). Darwin and emotion expression. American Psychological Association, 64(2): 120-128.

[2] Palaz, D., Doss, M.M., Collobert, R. (2015). Raw speech signal-based continuous speech recognition using convolutional neural networks. 2015 IEEE International

Conference on Acoustics, Speech and Signal Processing (ICASSP), Brisbane, QLD, pp. 4295-4299. https://doi.org/10.1109/ICASSP.2015.7178781

[3] Matthew, E.P., Waleed, A., Chandra, B., Russell, P. (2017). Semi-supervised sequence tagging with bidirectional language models. In ACL. pp. 1-10. https://arxiv.org/abs/1705.00108

[4] Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R. (2018). Glue: A multi-task bench mark and analysis platform for natural language understanding. Proc of ICLR Conference, USA, pp. 1-20. https://arxiv.org/abs/1804.07461

[5] Basharirad, B., Moradhaseli, M. (2017). Speech emotion recognition methods: A literature review. Proc of 2nd International Conference on Applied Science and Technology (ICAST'17) AIP, 1891(1): 020105. https://doi.org/10.1063/1.5005438

[6] Tamazin, M., Gouda, A., Khedr, M. (2019). Enhanced automatic speech recognition system based on enhancing power-normalized cepstral coefficients. Applied Sciences, 9(10): 2166. https://doi.org/10.3390/app9102166

[7] Kubanek, M., Bobulski, J., Kulawik, J. (2019). A method of speech coding for speech recognition using a convolutional neural network. Symmetry, 11(9): 1185. https://doi.org/10.3390/sym11091185

[8] Nagajyothi, D., Siddaiah, P. (2018). Speech recognition using convolutional neural networks. International Journal of Engineering and Technology, 7(4.6). https://doi.org/10.14419/ijet.v7i4.6.20449

[9] Abdel-Hamid, O., Mohamed, A.R., Jiang, H., Deng, L., Penn, G., Yu, D. (2014). Convolutional neural networks for speech recognition. EEE/ACM Transactions on Audio, Speech, and Language Processing, 22(10): 1533-1545. https://doi.org/10.1109/TASLP.2014.2339736

[10] Cao, H., Verma, R., Nenkova, A. (2015). Speaker-sensitive emotion recognition via ranking: Studies on acted and spontaneous speech. Computer Speech & Language, 29(1): 186-202. https://doi.org/10.1016/j.csl.2014.01.003

[11] Nwe, T.L., Foo, S.W., Silva, L.C. (2003). Speech emotion recognition using hidden Markov models. Speech Communication, 41(4): 603-623. https://doi.org/10.1016/S0167-6393(03)00099-2

[12] Rong, J., Li, G., Chen, Y.P. (2019). Acoustic feature selection for automatic emotion recognition from speech. Information Processing and Management, 45(3): 315-328. https://doi.org/10.1016/j.ipm.2008.09.003

[13] Lee, C.M., Narayanan, S.S. (2005). Toward detecting emotions in spoken dialogs. IEEE Transactions on Speech and Audio Processing, 13(2): 293-303. https://doi.org/10.1109/TSA.2004.838534

[14] Yang, B., Lugger, M. (2010). Emotion recognition from speech signals using new harmony features. Signal Processing, 90(5): 1415-1423. https://doi.org/10.1016/j.sigpro.2009.09.009

[15] Albornoz, E.M., Milone, D.H., Rufiner, H.L. (2011). Spoken emotion recognition using hierarchical classifiers. Computer Speech & Language, 25(3): 556-570. https://doi.org/10.1016/j.csl.2010.10.001

[16] Lee, C.C., Mower, E., Busso, C., Lee, S., Narayanan, S. (2011). Emotion recognition using a hierarchical binary decision tree approach. Speech Communication, 53(9-10): 1162-1171. https://doi.org/10.1016/j.specom.2011.06.004

[17] Dai, W., Han, D., Dai, Y., Xu, D. (2015). Emotion recognition and affective computing on vocal social media. Information and Management, 52(7): 777-788. https://doi.org/10.1016/j.im.2015.02.003

[18] Suksri, S. (2012). Speech recognition using MFCC. International Conference on Computer Graphics, Simulation and Modeling (ICGSM'2012), Pattaya, Thailand, pp. 135-138. https://doi.org/10.13140/RG.2.1.2598.3208

[19] Vuppala, A.K., Rao, K.S., Chakrabarti, S. (2012). Improved consonant-vowel recognition for low bit-rate coded speech. International Journal of Adapt Control and Signal Processing, 26(4). https://doi.org/10.1002/acs.1286

[20] Benesty, J., Sondhi, M.M., Huang, Y. (2008). Springer Handbook of Speech Processing. Springer. pp. 1-400.

[21] Nassif, A.B., Shahin, I., Attili, I., Azzeh, M., Shaalan, K. (2019). Speech recognition using deep neural networks: A systematic review. IEEE Access, 7: 19143-19165. https://doi.org/10.1109/ACCESS.2019.2896880

[22] Ashar, A., Shahid, M., Mushtaq, U. (2020). Bhatti speaker identification using a hybrid CNN-MFCC approach. IEEE Conference, Karachi, Pakistan, pp. 1-6.

[23] Charan, R., Manisha, A., Karthik, R., Kumar, R. (2017). A text-independent speaker verification model: A comparative analysis. 2017 IEEE International Conference on Intelligent Computing and Control (I2C2), At India pp, 1-6. https://doi.org/10.1109/I2C2.2017.8321794

[24] Wang, X., Paliwal, K.K. (2003). Feature extraction and dimensionality reduction algorithms and their applications in vowel recognition. Pattern Recognition, 36(10): 2429-2439. https://doi.org/10.1016/S0031-3203(03)00044-X

[25] Moussa, S., Hajaiej, Z., Garsallah, A. (2019). Enhancement of speech signal denoising based on MFCC and robust principal component analysis RPCA. IJCSNS International Journal of Computer Science and Network Security, 19(3): 1-14.

[26] Cutajar, M., Gatt, E., Grech, I., Casha, O., Micallef, J. (2012). Comparative study of automatic speech recognition techniques. IET Signal Processing, 7(1): 25-46. https://doi.org/10.1049/iet-spr.2012.0151

[27] Rapuano, S., Harris, F. (2007). An introduction to FFT and time domain windows. IEEE Instrumentation and Measurement Magazine, 10(6): 32-44. https://doi.org/10.1109/MIM.2007.4428580

[28] Alim, S.A., Rashid, N.K. (2018). Some commonly used speech feature extraction algorithms. Natural to Artificial Intelligence - Algorithms and Applications, Ricardo Lopez-Ruiz, IntechOpen, pp. 1-12. https://doi.org/10.5772/intechopen.80419

[29] Gadekar, P., Kaldane, M.H., Pawar, D., Jadhav, O., Patil, A. (2019). Analysis of speech recognition techniques. Proc of International Journal of Advance Research, Ideas and Innovations in Technology, 5(2): 1-15.