Indonesian Journal of Electrical Engineering and Informatics (IJEEI)
Vol. 6, No. 1, March 2018, pp. 12~20
ISSN: 2089-3272, DOI: 10.11591/ijeei.v6i1.409

A Review on Emotion Recognition Algorithms Using Speech Analysis

Teddy Surya Gunawan*[1], Muhammad Fahreza Alghifari[2], Malik Arman Morshidi[3], Mira Kartiwi[4]
[1,2,3]Department of Electrical and Computer Engineering, International Islamic University Malaysia
[4]Department of Information Systems, International Islamic University Malaysia

Article Info

Article history:

Received Nov 15, 2017
Revised Dec 7, 2017
Accepted Dec 24, 2017

Keyword:

Deep neural networks
Emotion database.
MFCC
Speech emotion recognition

ABSTRACT

In recent years, there is a growing interest in speech emotion recognition (SER) by analyzing input speech. SER can be considered as simply pattern recognition task which includes features extraction, classifier, and speech emotion database. The objective of this paper is to provide a comprehensive review on various literature available on SER. Several audio features are available, including linear predictive coding coefficients (LPCC), Mel-frequency cepstral coefficients (MFCC), and Teager energy based features. While for classifier, many algorithms are available including hidden Markov model (HMM), Gaussian mixture model (GMM), vector quantization (VQ), artificial neural networks (ANN), and deep neural networks (DNN). In this paper, we also reviewed various speech emotion database. Finally, recent related works on SER using DNN will be discussed.

Corresponding Author:

Teddy Surya Gunawan,
Department of Electrical and Computer Engineering,
International Islamic University Malaysia
Jalan Gombak, 53100 Kuala Lumpur, Malaysia
Email: tsgunawan@iium.edu.my, tsgunawan@gmail.com

1. INTRODUCTION

Speech emotion recognition (SER) is one of the topics in speech processing that has been continuously researched. The initial start of that is simple speech recognition dates back from the late fifties [1]. In today's world, SER has shown to be quite a research hotspot, as indicated by the growth of publication papers in each year. Figure 1 shows the rough estimation of IEEE published papers that are related to SER. The data was analyzed from IEEE Explore.

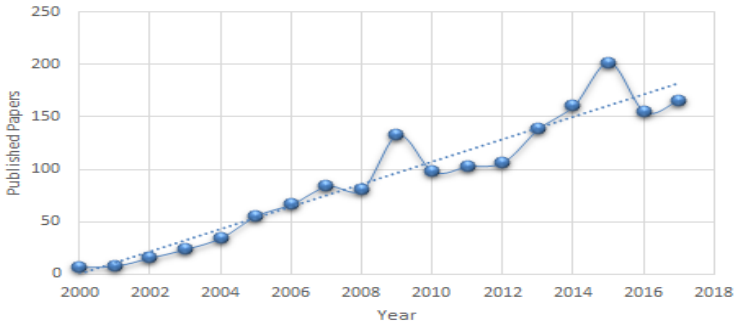

Figure 1. Growth of IEEE SER Published Papers from 2000 until 2017

The aim of SER system is to extract the emotion from the unknown input speeh [2]. While each individual may have their own abstract emotional state, generally emotions can be grouped into a universal category of happiness, anger, surprise, fear, sadness as well as neutral. Some other researchers have their own categories, for example the database utilized in [3] categorized emotions into ten types, namely joy, acceptance, fear, surprise, sadness, disgust, anger, anticipation, neutral, and others. Although the classification of emotion might differ, the objective of SER is still the same, which is to extract emotional state. In [4], it is stated that SER is more or less a pattern recognition system. Figure 2 shows the typical speech emotion recognition (SER) system.

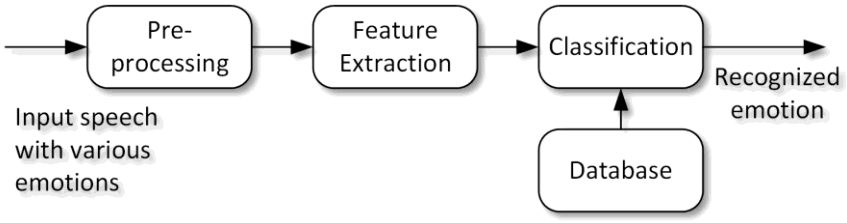

Figure 2. Typical Speech Emotion Recognition System

The application of SER can be targeted to several sectors. In banking, an auto caller equipped with SER may assist in detecting the emotion of the customer, generating custom responses based on the result [5-7]. In education, an e-learning portal with SER can detect the emotions of the user such as frustration and stress, determining whether the studying is conducive or not and give appropriate countermeasures [8]. Yet another application is in transportation, where in the near-future that vehicles are capable of auto-driving, the system can take over the steering wheel in the case where an unhealthy amount of emotion is detected from the driver [9].

2. REVIEW ON AUDIO FEATURES EXTRACTION

In this section, various audio features used in SER are reviewed, including linear predictive coding coefficients (LPCC), Mel-Frequency Cepstral Coefficients (MFCC), and Teager energy operator (TEO). The extraction process goes through three steps. First, the pre-emphasis is a filter used to emphasize on high frequency baud by increasing its amplitude and decreasing the amplitude of lower frequency. In speech, typically the higher frequency holds more important information to extract, while lower frequency might be mingled with noise. It should be noted that in modern speech recognition systems the pre-emphasis has lost its importance and replaced by channel normalization in the later steps, but for the sake of simple but effective methods, a high-pass filter is sufficient.

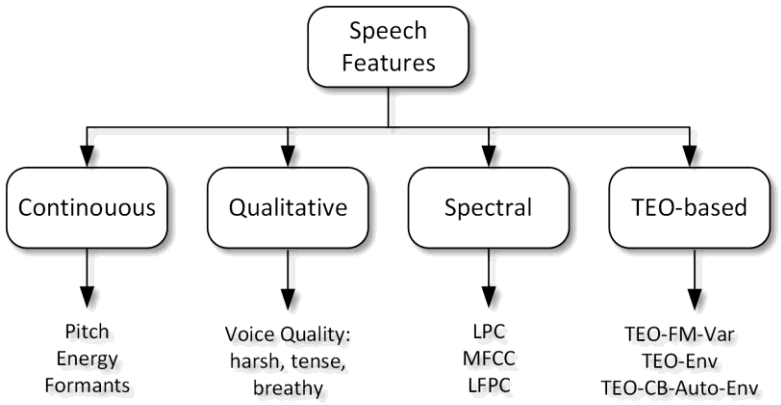

Figure 3. Categories of Speech Features [1]

Secondly, the frame blocking and windowing is a process to decompose the speech signal into short speech sequences called frames to conduct speech analysis. There are several windows that can be utilized such as the rectangle window, triangular window, but the Hamming window is often chosen as it softens the edges created due to framing, again emphasizing on simplicity. Third is the feature extraction. According to [1], speech features can be categorized into four groups, including namely continuous, qualitative, spectral, and TEO-based features, as shown in Figure 3.

2.1 Linear Predictive Coding Coefficients (LPCC)

Linear predictive coding (LPC) is a digital method for encoding an analog signal [10]. The way LPC works is that it predicts the next value of a signal based on the information it has received in the past, forming a linear pattern. The main objective of LPC to obtain a set of predictor coefficients that will minimize the mean squared error, E_m. The formula used to obtain the LPC coefficients is:

$$E_m = \sum_n e_m^2[n] = \sum_n \left(x_m[n] - \sum_{j=1}^{p} a_j x_m[n-j] \right)^2 \quad (1)$$

where $x_m[n]$ is a frame of the speech signal and p the order of the LPC analysis. LPC encoding generally gives satisfactory quality speech at a lower bit rate and supplies pinpoint approximations of speech parameters. Although LPCC can be considered one of the more traditional features of speech, LPC has contribute to the overall recognition of emotion. In [11], they used LPCC as one of their features and achieved 86.41% recognition.

2.2 Mel-Frequency Cepstral Coefficients (MFCC)

The Mel-frequency cepstral coefficients (MFCC) is one of the most popular audio feature [12, 13]. It is a representation of the speech signals where a feature called the cepstrum of a windowed short-time signal is derived from the FFT of that signal. Afterwards the signal goes to the frequency axis of the mel-frequency scale using a log based transform, and then decorrelated using a modified Discrete Cosine Transform [14].

The steps to extract MFCC features, including pre-emphasis, frame blocking and windowing, FFT magnitude, Mel filterbank, log energy, and DCT as explained in [13]. MFCC utilizes the mel-scale, which is tuned to the human's ear frequency response. Due to this, MFCC has been proven to be invaluable in the speech recognition field, and has been attempted to be integrated with emotion recognition [15]. According to [1], Spectral audio features such as MFCC is best suited for a N-way classifiers.

2.3 Eager Energy Operator (TEO)

The Teager Energy Operator (TEO) was proposed by Herbert M. Teager and Shushan M. Teager in 1983. In their article, they argued that the speech model at that time was inaccurate due to its linear finite characteristics, and proposed a model that involves a nonlinear process. Later in another article, they generated a plot that implies the energy creating the sound, but the algorithm was not specified [16]. The works is further extended in [17] and Teager Energy Operator has since been defined for both real and complex continuous signals. TEO can be defined as

$$\psi\big((x)t\big) = \dot{x}^2(t) - x(t)\ddot{x}(t) \quad (2)$$

TEO has been used in various speech signal applications. In [16], formants of vowels are tracked using TEOs. In SER, TEO features are used by [18] to make their system more robust in noisy environment. Moreover, TEO-based features are suitable to detect the stress level of emotion [1].

2.4 Summary of Various Audio Features

The features to be extracted are various, but they can be grouped into 4 distinct groups, namely continuous, qualitative, spectral, and TEO-based features. These features can be used as a sole determinant, but often they are used in combination to generate a more distinguishable pattern for the system. Table 1 shows the strength and weaknesses of various audio features. We selected MFCC due to its suitability for N-based classifiers and DNN. Moreover, many researches have used MFCC as the audio features. So that, our proposed system could be benchmarked with other research.

Table 1. Summary of literature review on audio features for SER

Audio Features	Strengths	Weaknesses
LPCC	One of the most traditional features which implies that it is widely recognized and used.	LPC on its own has is not as reliable, as seen that it is often combined with other feature extraction methods.
MFCC	Tuned in a scale that is suitable for the human ear. Alongside with LPCC, is considered one of the standard features extracted, even more-so in SER. Best suited for N-way classifiers.	MFCC being in spectral form is sensitive towards noise.
TEO	Nonlinear approach, which is for suitable for speech. Superior detection in stress-levels of emotion.	More complicated computations as compared to LPC.

3. REVIEW ON CLASSIFIERS

After the SER system extracts the desired features from the audio speech data, the next step is to pass the data on to the classifier. The primary job of the classifier is to determine the unrevealed emotion of the user by using a set of defined algorithms and functions. Usually these classifier evaluations are performed using a single database or dataset, under one language. Up until now, there has been no agreed standard of which classifier is the best, but many have been evaluated to achieved better recognition. The ones that are most commonly used classifier are: GMMs, HMMs, SVMs ANNs as well as k-NN [1]. In this section, the three most popular classifiers HMM, GMM and VQ are discussed in brief and compared with the classifier that is used in this project, Deep Neural Network DNN, which is an extended version of ANN.

3.1 Hidden Markov Model (HMM)

The Hidden Markov Model (HMM) consist of the first order markov chain whose states are hidden from the observer. This means while that the observer cannot directly examine the internal behavior of the model as it remains hidden, the the data's temporal structure is recorded by these states. HMM can be considered as statistical models that describe the sequences of events [2]. To express this in mathematical terms, for modeling a sequence of observable data vectors, $x_1, \cdots, x_T$ by an HMM, we assume the existence of a hidden Markov chain responsible for generating this observable data sequence. Let Kbe the number of states, $\pi_i, i = 1, \cdots, K$ be the initial state probabilities for the hidden Markov chain, and $a_{ij}, i = 1, \cdots, K, j = 1, \cdots, K$ be the transition probability from state I to state j. Assuming the true state sequence is $s_1, \cdots, s_T$ the likelihood of the observable data is given by

$$p(x_1, s_1, \cdots, x_T, s_T) = \pi_{s_1} b s_1(x_1) a_{s_1 s_2} b_{s_2}(x_2) \ldots a_{s_{T-1} s_T} b_{s_T}(x_T)$$
$$p(x_1, s_1, \cdots, x_T, s_T) = \pi_{s_1} b_{s_1}(x_1) \prod_{t=2}^{T} a_{s_{t-1} s_t} b_{s_t}(x_t) \quad (3)$$

HMM is also a sequential generating probabilistic model, which means that the classifier acts on the assumption that neighboring frames are closely related. While this is valid for speech signal frames, there are better alternatives due to its assumption and algorithm complexity [19].

3.2 Gaussian Mixture Models (GMM)

The Gaussian mixture model (GMM) uses alternate generating probabilistic model, which implies that for a particular word we can form multivariate Gaussian density models that represents all the frames [19]. Similar to HMM, GMM can be expressed in mathematical terms. Let $P(x_t)$ be the t-th frame of the isolated word x. The probability of generating the frame Let $P(x_t)$ using GMM can computed as follows:

$$P_{GMM}(x_t) = \sum_{k=1}^{s} c_k G_k(x_t) \quad (4)$$

where s is the number of mixtures, c_k is the probability of the k^{th} mixture, and G_k is the multivariate Gaussian density function with mean vector and covariance matrix. Compared to HMMs, GMM are superior in training and testing due to their efficiency in modeling multi-modal distributions as a whole. GMMs are used in SER when global features are the main focus. But due to this feature, GMMs are not suited when the user would like to model the temporal structure.

3.3 Vector Quantization (VQ)

Vector quantization (VQ) is a process of mapping feature vectors of test utterance to the best matching feature vectors of the reference models [20]. As compared to other techniques such as HMM, VQ boosts is its low computational burden due to its straightforward approach. The efficiency is due to its nature of using compact codebooks for reference models and codebook searcher [21]. While the basic VQ appears to be convenient, because the vectors are jumbled up, VQ does not take into account the temporal evolution of the signals.

3.4 Artificial Neural Network (ANN) and Deep Neural Network (DNN)

The term artificial neuron network (ANN) is a term commonly used for a system that imitates the flow of the neuron. Information is received from the input and flows from one node to another, until it reaches the output. Through this process, the system will learn about the input given. Three branches of ANN will be discussed, including feedforward neural network, deep neural network and convolutional neural network.

The feedforward neural network is the first type of neural network developed. The process is the most basic one of all: the data is forwarded through an input layer to a single hidden layer, then to the output layer. In feedfoward, there are no loops or cycles. In the feedforward neural network, there is the input layer, a hidden layer, and an output layer. A deep neural network expands the possibilities by adding more layers in the hidden layer segment [22] [23]. An interesting characteristic of DNNs is that they can learn high-level invariant features from raw data.

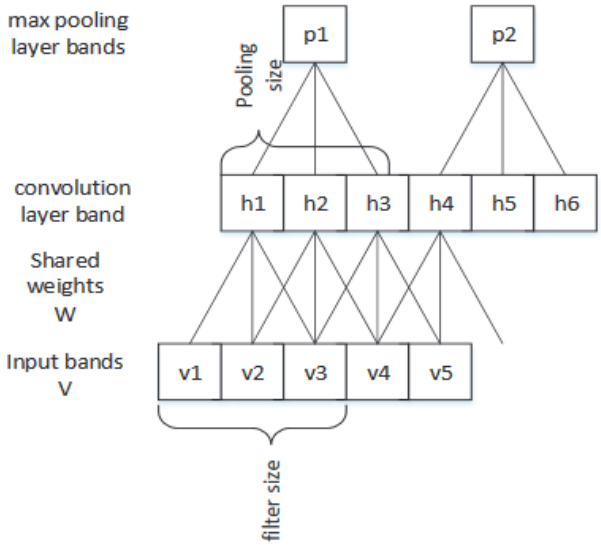

Figure 4. Convolutional Neural Network [24]

Convolutional neural network (CNN), as shown in Fig. 4, is inspired by the visual cortex, where cells are activated according to their sub-regions. Applying that to ANN, the CNN information in the neurons are connected to their sub-regions first, before passing to the next layer. Some sub-regions may overlap. This contrasts with other neural network architectures where each neuron is independent [25]. While CNNs are highly sophisticated and can be used for SER, it is specifically suitable for image processing and recognition, due to the convolutional layer.

The classifier is the algorithm that determines how these features are manipulated and translated into emotion recognition. Common classifiers are HMM, SVM, GMM, and ANN. DNN is used, a more sophisticated version of feedforward ANN. Table 2 shows that ANN boosts deep potential for pattern recognition, provided that more layers are supplied. The weakness of inconvenience when adding emotion can be simply solved by consolidating all initial parameters at the start. This claim is further supported in Table 3, where using DNN may generate more accurate recognition compared to other classifiers.

Table 2. Summary of literature review on classifier

Classifier	Advantages	Disadvantages
HMM	• Text-independent	• Significant increase in computational complexity • The need of a proper initialization for the model parameters before training • comparatively takes time to train
VQ	• Unsupervised clustering • Low computational burden • Text independent	• Text-dependent • Does not take temporal evolutions into account.
GMM	• probabilistic framework (robust) • computationally efficient • easy to be implemented • Suited for extracting global features	• Require to gather all the model parameters before starting • comparatively takes time to train
ANN	• Less parameters • Higher performance compared to VQ model • Has very high potential given more hidden layers.	• Network must be retrained when a new emotion is added to the system

4. REVIEW ON SPEECH EMOTION DATABASE

To complete the process of SER, the system requires a database for training and testing. An emotion database generally consists of various audio recordings that are labeled their appropriate emotion. For this section, the discussion will be directed towards the number of databases used, the method of obtaining the dataset, the variety of emotions categorized, as well as the challenges that most researches have in obtaining these databases.

Usually a single SER system will rely only on a single database, to reduce data variance due to external factors such as different accents. While most systems are supported by one database, there are some researches that utilizes more, such as by [26], that have used the Berlin emotional speech database (EMO-DB) in combination with the German FAU Aibo emotion corpus (FAUAEC). With that said, these databases are still only using one language; German. As previously mentioned, there are external factors that can affect the speech features that are extracted.

The closest attempt of integrating multiple databases was performed by [27] by using 6 standard databases (AVIC, DES, EMO-DB, eNTERFACE, SmartKom, SUSAS) in a cross-corpora and multilingual evaluation experiment. An alternative is using a database that has already integrated multiple languages, such as the INTERFACE corpus, which supports English, Slovenian, Spanish, and French.

Another aspect to consider is *how* these speech emotion data are obtained. One may debate that true authentic emotion can only be captured at the moment, but spontaneous speech is difficult to record. To ensure proper speech processing, the system requires better audio quality. This is simply not feasible to attain without proper sound recording setup and environment. Therefore, the most used method is for professional or experienced actors to express the emotion through acting, then labeling each speech segment on its appropriate category. The EMO-DB and LDC Emotional Prosody Speech and Transcripts are two examples of an actor-based database. Generally, this is conducted under ideal conditions (ie: in a studio with minimum noise interference).

Another interesting method of collecting data is by collecting the speech from existing media, such as from movies, television recording, etc. While the source can be still considered as a "professional actor", the method of collecting the data differs from the first but maintains the general quality of audio. This however is met with the problem of copyright of fair usage. An example of a research that utilizes this method is by [28]. Finally, there are researches that collect their data from non-professional actors. These databases are generally self-made from the local environment. But while a home-made database creation may be more convenient for the researcher, it becomes difficult to benchmark the results with other papers.

There are variations of emotions that are categorized. The German Database for example, groups the emotion into anger, boredom, disgust, fear, happy, neutral, and sad. The more emotions category the database has, the more challenging it is for the SER system to achieve high accuracy. To solve this, some researches such as [3] merges and omits certain emotions with similar attributes, eg, the emotion of "disgust" and "anger". and focuses on those emotions with distinct variations.

There are various other factors to consider when choosing the appropriate database such as number of actors, language, ethnicity, word utterance or whole sentence, but one factor that has been a deterrent for some young researchers is the fact that some databases are obscured by a pay wall. This leads to either creating their own database or using open-source databases available. Table 3 shows various databases along with the audio features and classifiers that are used by other researchers.

Table 3. Summary of Recent Researches on Audio Features, Classifiers, and Databases

Ref	Acoustic Features	Classifiers	Emotions	Databases
[29]	LPCC, MFCC, pitch, intensity and formant	SVM	Neutral, angry and sad	Self built Bengali emotional database (actor based)
[30]	Micro perturbations in pitch (jitter) & very small variations in intensity (shimmer).	ANN	Happy, surprise, neutral, anger, sad, fear, disgust	Self built English Hindi database (induced) 2765 wave files in English and 2240 wave files in Hindi
[31]	Error correcting codes (ECC)	ANN	Anger, boredom, disgust, fear, happy, neutral, sad	Emo-DB database
[11]	MFCC, Perceptual Linear Prediction cepstral coefficient (PLP), LPCC	SVM	Anger, fear, happiness, neutral, surprise, sadness	Chinese Academy of Sciences Institute of automation of speech emotion database (CASIA)
[15]	Pitch, intensity, speech rate, MFCC	SVM	Anger, fear, happiness, sad, neutral	Toronto emotional speech set (TESS) collection
[32]	prosodic features: energy contour and pitch contour spectral features: cepstral coefficients, quefrency coefficients and formant frequencies	Hybrid Rule based K-mean clustering, SVM	Anger, happy, sad	Amritaemo (local database)
[26]	MFCC, log energy, delta and acceleration coefficients	Kernel sparse representation based classifier (KSRC)	EMO-DB: disgust, sadness, fear, happiness, boredom, neutral, anger. FAUAEC: Anger, emphatic, neutral, motherese	The Berlin emotional speech database (EMO-DB) and The German FAU Aibo emotion corpus (FAUAEC)
[3]	TPCC (Teager Phase Cepstrum Coefficients), RPCC (Residual Phase Cepstrum Coefficients), and GLFCC (Glottal Flow Cepstrum Coefficients)	SVM	Joy, fear, sadness, anger, neutral	Online gaming voice chat corpus with emotional labels (OGVC)
[33]	MFCCs, Teager energy operator (TEO) and glottal time and frequency domain parameters.	GMM	Anger, boredom, disgust, fear, happiness, neutral, sadness	Berlin Emotional Speech database
[34]	Pitch	HMM, SVM	Anger, happiness, sadness and neutral	No database specified
[35]	Fusion of TEO and MFCC	GMM	Angry, anxiety, disgust, neurtral, sadness	Berlin Emotional Speech database
[36]	Pitch, loudness & formant	GMM	joy, anger, surprise and sadness	Self built emotional database in 60 different statements
[37]	Spectogram	Convolutional DNN	Anger, boredom, disgust, fear, happiness, neutral, sadness	Berlin Emotional Speech database
[38]	Energy, pitch, voice probability, and 26-dimensional log Mel-spectrogram features	GMM, ELM, DNN	Neutral, happy, sad, angry.	10527 real-traffic Mandarin usable utterances from a Microsoft spoken dialogue system.

5. RELATED WORKS AND PROPOSED SER SYSTEM

Fortunately, SER is a topic that is abundant in papers these recent years. In 2017 only, there are more than 150 papers published that are related to SER, which covers different angles of approaches, new combination of features to be processed, implementation of a variety of algorithms, optimization of results. A brief sample of research methodologies conducted in 10 papers can be observed in Table 3. Table 4 shows additional closely related papers, i.e. SRR using DNN. Although many researches have been conducted on SER using various audio features, classifiers, or database, however, there is still a need to further improve the accuracy and processing time of an SER system.

Table 4. Summary of Related Works on SER using DNN

Source	Methodology	Strengths	Weaknesses
Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine [39]	MFCC features, pitch-based features -pitch period and harmonics-to-noise ratio (HNR), their delta feature across time frames. Interactive Emotional Dyadic Motion Capture (IEMOCAP) database excitement, frustration, happiness, neutral and surprise	Well described methodology on how to conduct experiment. Results show that proposed DNN methodology outperforms HMM and SVM by 20% relative accuracy. ELMs paradigm proposed are 10 times faster than SVMs.	While comparison analysis between DNN and HMM and SVM is attempted, less information is supplied on how the latter (HMM & SVM) is performed. While weighted and unweighted average is more accurate, the overall recognition rate is not mentioned.
Acoustic Emotion Recognition using Deep Neural Network [40]	MFCCs, perceptual linear predictive (PLPs), and Filter banks (FBANKs) 9595 emotion sentences (no named database) Angry, happy, fear, sad, surprise, neutral.	Better documentation of comparison analysis between GMM and DNN, under equal conditions. DNN accuracy shows 8.2 percentage points increase compared with baselines GMMs, up to 92.3%.	No mention of processing time. No mention of database for the sake of verifying results or comparisons.
Deep Learning Based Affective Model for Speech Emotion Recognition [23]	Feature extraction is managed automatically by deep networks. German Berlin Emotional Speech Database anger, boredom, disgust, anxiety, happiness, sadness and neutral state	Proposes a affective system that will choose the features by itself. Recognition accuracy reaches 65% in the best case, an improvement from their benchmark, which is 22%.	Using a system that chooses the appropriate features is promising, but there should be mentions of which distinct features are prominent. Lack of depth implies dependency on toolbox.
Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network [37]	Features extracted from spectrogram generated from speech. German Berlin Emotional Speech Database anger, boredom, disgust, anxiety, happiness, sadness and neutral state	Uses novel approach of using image recognition of spectrogram generated from speech. Achieved a recognition rate of 84.3%.	The need to analze the spectrogram of the audio adds a layer of complexitiy to the SER system, which may not be applied in real life usge.
Speech emotion recognition based on Gaussian Mixture Models and Deep Neural Networks [38]	Energy, pitch, voice probability, and 26-dimensional log Mel-spectrogram features, total 58 features from each frame. Total 10; 527 real-traffic Mandarin usable utterances from a Microsoft spoken dialogue system. Neutral, happy, sad, angry.	Provides reliable and comprehensive documentation of data collection. 4 different algorithms are applied: GMM, DNN, and 2 variation of extreme machine learning (ELM).	With the large amount of emotional utterance, more variation of emotion classification should be possible. While it leaves more room for future researches to improve, the best recognition rate is only 57.9% using ELM-DNN,

Based on Table 4, we proposed SER system as shown in Figure 5. The raw audio received from the EMO-DB is labeled into their respective emotions. These audios are then inserted into a temporary storage for feature extraction. The next step is feature extraction using MFCC. Finally, the extracted features are classified using DNN. The performance evaluation of the proposed system will be discussed in our next paper.

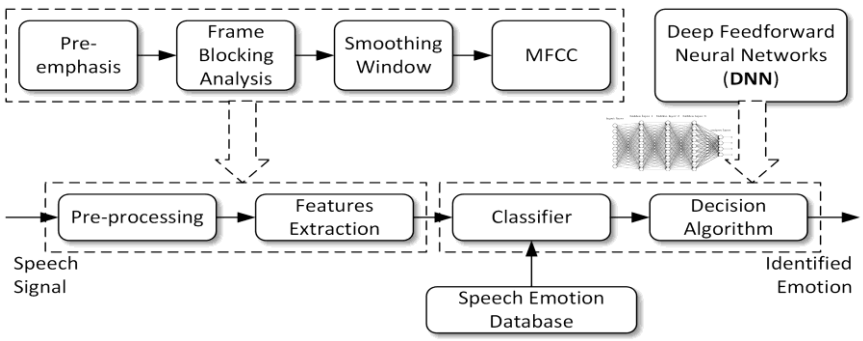

Figure 5. Proposed SER System

6. CONCLUSION

This paper has presented a comprehensive review on the emotion recognition using speech analysis and the design of SER system. A typical SER consisted of at least feature extraction, classifier, and speech emotion database. From the critical literature review, of the various audio features we selected MFCC due to its popularity and suitability, while deep neural network was selected as the classifier due to its higher accuracy if more data is available. A comprehensive and popular emotion database, EMO-DB, was selected. Further research includes implementation of the proposed SER system using Matlab and performance evaluation and benchmarking.

ACKNOWLEDGEMENT

The authors would like to express their gratitude to the Malaysian Ministry of Higher Education (MOHE), which has provided funding for the research through the Research Acculturation Grant Scheme, RAGS15-070-0133.

REFERENCES

[1] M. El Ayadi, M. S. Kamel, and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases," *Pattern Recognition*, vol. 44, pp. 572-587, 2011.
[2] A. Joshi and R. Kaur, "A Study of speech emotion recognition methods," *Int. J. Comput. Sci. Mob. Comput.(IJCSMC)*, vol. 2, pp. 28-31, 2013.
[3] M. Takebe, K. Yamamoto, and S. Nakagawa, "Investigation of glottal features and annotation procedures for speech emotion recognition," in 2016 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA), pp. 1-4, 2016.
[4] A. B. Ingale and D. S. Chaudhari, "Speech emotion recognition," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, pp. 235-238, 2012.
[5] D. Pappas, I. Androutsopoulos, and H. Papageorgiou, "Anger detection in call center dialogues," in Cognitive Infocommunications (CogInfoCom), 2015 6th IEEE International Conference on, pp. 139-144, 2015.
[6] J. Irastorza and M. I. Torres, "Analyzing the expression of annoyance during phone calls to complaint services," in Cognitive Infocommunications (CogInfoCom), 2016 7th IEEE International Conference on, pp. 000103-000106, 2016.
[7] Y. Guo, Y. Li, Q. Wei, and S. X. Xu, "IT-Enabled Role Playing in Service Encounter: Design a Customer Emotion Management System in Call Centers," 2017.
[8] L. Kerkeni, Y. Serrestou, M. Mbarki, K. Raoof, and M. A. Mahjoub, "A review on speech emotion recognition: Case of pedagogical interaction in classroom," in Advanced Technologies for Signal and Image Processing (ATSIP), 2017 International Conference on, pp. 1-7, 2017.
[9] J. S. K. Ooi, S. A. Ahmad, H. R. Harun, Y. Z. Chong, and S. H. M. Ali, "A conceptual emotion recognition framework: stress and anger analysis for car accidents," *International journal of vehicle safety*, vol. 9, pp. 181-195, 2017.
[10] A. Dixit, A. Vidwans, and P. Sharma, "Improved MFCC and LPC algorithm for bundelkhandi isolated digit speech recognition," in 2016 International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT), pp. 3755-3759, 2016.
[11] W. Fei, X. Ye, S. Zhaoyu, H. Yujia, Z. Xing, and S. Shengxing, "Research on speech emotion recognition based on deep auto-encoder," in 2016 IEEE International Conference on Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), pp. 308-312, 2016.
[12] T. S. Gunawan, N. A. M. Saleh, and M. Kartiwi, "Development of Quranic Reciter Identification System using MFCC and GMM Classifier," *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 8, 2018.
[13] T. S. Gunawan and M. Kartiwi, "On the Comparison of Line Spectral Frequencies and Mel-Frequency Cepstral Coefficients Using Feedforward Neural Network for Language Identification," *Indonesian Journal of Electrical Engineering and Computer Science*, vol. 10, 2018.
[14] X. Huang, A. Acero, and H.-W. Hon, *Spoken language processing: A guide to theory, algorithm, and system development*, Prentice hall PTR, 2001.
[15] D. Verma and D. Mukhopadhyay, "Age driven automatic speech emotion recognition system," in 2016 International Conference on Computing, Communication and Automation (ICCCA), pp. 1005-1010, 2016.
[16] E. Kvedalen, "Signal processing using the Teager energy operator and other nonlinear operators," *Master, University of Oslo Department of Informatics*, vol. 21, 2003.
[17] J. F. Kaiser, "On a simple algorithm to calculate the'energy'of a signal," pp. 381-384, 1990.
[18] A. Georgogiannis and V. Digalakis, "Speech Emotion Recognition using non-linear Teager energy based features in noisy environments," in 2012 Proceedings of the 20th European Signal Processing Conference (EUSIPCO), pp. 2045-2049, 2012.
[19] E. Gopi, *Digital speech processing using Matlab*, Springer, 2014.

[20] M. Saleem, Z. U. Rehman, U. Zahoor, A. Mazhar, and M. R. Anjum, "Self learning speech recognition model using vector quantization," in 2016 Sixth International Conference on Innovative Computing Technology (INTECH), pp. 199-203, 2016.
[21] M. A. Anusuya and S. K. Katti, "Speech recognition by machine, a review," *arXiv preprint arXiv:1001.2267*, 2010.
[22] D. Yu and L. Deng, *Automatic speech recognition: A deep learning approach*, Springer, 2014.
[23] X. Zhou, J. Guo, and R. Bie, "Deep Learning Based Affective Model for Speech Emotion Recognition," in 2016 Intl IEEE Conferences on Ubiquitous Intelligence & Computing, Advanced and Trusted Computing, Scalable Computing and Communications, Cloud and Big Data Computing, Internet of People, and Smart World Congress (UIC/ATC/ScalCom/CBDCom/IoP/SmartWorld), pp. 841-846, 2016.
[24] D. Guiming, W. Xia, W. Guangyan, Z. Yan, and L. Dan, "Speech recognition based on convolutional neural networks," in 2016 IEEE International Conference on Signal and Image Processing (ICSIP), pp. 708-711, 2016.
[25] M. H. Beale, M. T. Hagan, and H. B. Demuth, "Neural network toolbox user's guide," pp., 2017.
[26] P. Sharma, V. Abrol, A. Sachdev, and A. D. Dileep, "Speech emotion recognition using kernel sparse representation based classifier," in 2016 24th European Signal Processing Conference (EUSIPCO), pp. 374-377, 2016.
[27] B. Schuller, B. Vlasenko, F. Eyben, M. Wollmer, A. Stuhlsatz, A. Wendemuth, and G. Rigoll, "Cross-Corpus Acoustic Emotion Recognition: Variances and Strategies," *IEEE Transactions on Affective Computing*, vol. 1, pp. 119-131, 2010.
[28] B. Schuller, S. Reiter, R. Muller, M. Al-Hames, M. Lang, and G. Rigoll, "Speaker Independent Speech Emotion Recognition by Ensemble Classification," in 2005 IEEE International Conference on Multimedia and Expo, pp. 864-867, 2005.
[29] A. Mohanta and U. Sharma, "Bengali speech emotion recognition," in 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), pp. 2812-2814, 2016.
[30] A. Jacob, "Speech emotion recognition based on minimal voice quality features," in 2016 International Conference on Communication and Signal Processing (ICCSP), pp. 0886-0890, 2016.
[31] R. Chakraborty and S. K. Kopparapu, "Improved speech emotion recognition using error correcting codes," in 2016 IEEE International Conference on Multimedia & Expo Workshops (ICMEW), pp. 1-6, 2016.
[32] S. S. Poorna, C. Y. Jeevitha, S. J. Nair, S. Santhosh, and G. J. Nair, "Emotion recognition using multi-parameter speech feature classification," in 2015 International Conference on Computers, Communications, and Systems (ICCCS), pp. 217-222, 2015.
[33] A. Albahri and M. Lech, "Effects of band reduction and coding on speech emotion recognition," in 2016 10th International Conference on Signal Processing and Communication Systems (ICSPCS), pp. 1-8, 2016.
[34] P. P. Ladde and V. S. Deshmukh, "Use of Multiple Classifier System for Gender Driven Speech Emotion Recognition," in 2015 International Conference on Computational Intelligence and Communication Networks (CICN), pp. 713-717, 2015.
[35] S. R. Bandela and T. K. Kumar, "Stressed speech emotion recognition using feature fusion of teager energy operator and MFCC," in 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), pp. 1-5, 2017.
[36] P. Patel, A. Chaudhari, R. Kale, and M. Pund, "Emotion recognition from speech with gaussian mixture models & via boosted gmm," *International Journal of Research In Science & Engineering*, vol. 3, 2017.
[37] A. M. Badshah, J. Ahmad, N. Rahim, and S. W. Baik, "Speech emotion recognition from spectrograms with deep convolutional neural network," in Platform Technology and Service (PlatCon), 2017 International Conference on, pp. 1-5, 2017.
[38] I. J. Tashev, Z.-Q. Wang, and K. Godin, "Speech Emotion Recognition based on Gaussian Mixture Models and Deep Neural Networks," in Information Theory and Applications Workshop (ITA), 2017, pp. 1-4, 2017.
[39] K. Han, D. Yu, and I. Tashev, "Speech emotion recognition using deep neural network and extreme learning machine," in Interspeech, pp. 223-227, 2014.
[40] J. Niu, Y. Qian, and K. Yu, "Acoustic emotion recognition using deep neural network," pp. 128-132, 2014.