Emotion Recognition by Facial Expression and Voice: Review and Analysis

Article *in* Journal of Informatics and Web Engineering · September 2022

DOI: 10.33093/jiwe.2022.1.2.4

4 authors, including:

K.-W. Ng
Multimedia University
73 PUBLICATIONS **252** CITATIONS
SEE PROFILE

Su-Cheng Haw
Multimedia University
172 PUBLICATIONS **814** CITATIONS
SEE PROFILE

Journal of Informatics and Web Engineering
Vol. 1 No. 2 (September 2022) eISSN: 2821-370X

Emotion Recognition by Facial Expression and Voice: Review and Analysis

Yixen Lim[1], Kok-Why Ng[2*], Palanichamy Naveen[3], Su-Cheng Haw[4]
[1,2,3,4] Multimedia University, Malaysia.
**Corresponding author: (kwng@mmu.edu.my)*

Abstract - Emotion is a scorching topic in the recent years due to the critical unseen stress incurred during the pandemic and post-pandemic. This is worsening with the recent economy's inflation and increase of living cost, many employees are seriously affected and drawn forth many families saddened cases and tremendous drop of working performance. The increasing stress brings a lot of harm not only to the individual but to the company's and country's growth. To recognize emotion through a single model is less accurate, however, recruiting multiple-models may lead to latency in data processing and possibly misleading results if the input models data are not properly filtered and segmented. This paper will review, analyze and theoretically compare 15 facial expression methods and 17 voice methods of emotion recognition research works. It will outline the pros and cons of each method and discuss the accuracy of some of the standalone and hybrid emotion recognition methods. Some of the methods (such as CNN, KNN and SVM) can span over multiple-models, but reveal different level of strengths. This is very important to discover, so that one may replace or enhance the weaker level if applying the same method across the multiple-models. This paper will also illustrate different levels of popularity of the methods in each model for visual comparison in ease. Hopefully, it can cater the new researchers a quick identification on the most suitable method for recognizing the emotion through facial expression and/or voice.

Keywords—Emotion Recognition, Facial Expression, Voice, Stress, Image Processing.

Received: 17 June 2022; Accepted: 12 September 2022; Published: 16 September 2022

I. INTRODUCTION

Emotion plays an important role in our daily lives, especially in the workplace. It can affect the performance of a company indirectly from different perspectives. There are many benefits if employees can maintain a positive emotion during their work. Positive emotions can boost creative thinking and interpersonal relationships, which leading the employees to have higher job satisfaction. However, there are too little time for the employers to observe and understand the emotions of their employees nowadays.

In this paper, we will discuss, analyze and compare theoretically the 15 methods of emotion recognition through facial expression and 17 methods of emotional recognition through voice. Readers can base on their need and resources to choose the best-match method for performing the emotion recognition event. Section 2 will discuss the

Journal of Informatics and Web Engineering
https://doi.org/10.33093/jiwe.2022.1.2.4

Published by MMU Press. URL: https://journals.mmupress.com/jiwe

15 methods of emotional recognition through facial expression. Section 3 will discuss the 17 methods of emotion recognition through voice. Section 4 will illustrate the popularity of the methods in the recent years.

II. LITERATURE REVIEW

A. Emotion recognition by facial expression

CNN is used to recognize the facial expression. VGGNet, AlexNet and Inception are some of the famous CNN architectures to classify images [1]. In CNN, there are many kernel sizes, such as 2, 4, 8, 16, 32, 64 and number of filters, such as 2, 4, 8, 16, 32, 64, 128, 256. Different size of kernel and filter numbers can lead to different result and accuracy. It is very important to select a suitable kernel size because low kernel sizes (such as 2) may lead to a network highly unstable. But, the authors stated that the kernel sizes such as 32 and 64 would not achieve convergence for some merging of parameters in the model; while moderate and suitable kernel sizes (such as from 8 to 16) would converge very well. One of the existing models which is not uniform and has different filters number across depth, is able to achieve a high accuracy of 65% that makes it performs the best for the Facial Expression Recognition 2013 (FER-2013) dataset.

Deep Belief Network (DBN) is further explored and discussed in [2]. DBN is obtained by training and stacking some layers of Restricted Boltzmann Machines (RBM) in a greedy manner. Some advantages of DBN are able to search for a suitable set of model parameters quickly and efficient in computation of latent variables in deepest layer. However, it also has some drawbacks. If the top-down influences on the inference process is ignored, the model trained might be unable to interpret some ambiguous sensory input. Also, it learns a layer of features at a time and did not carry out the readjustment of the parameters at lower level. It is also slow and inefficient.

Some interesting and effective feature extraction methods are discussed in [3]. Haralick texture features that is obtained from the Gray-Level Co-occurrence Matrix (GLCM). The texture information will emerge from the features obtained. Next, Local Binary Pattern (LBP) is used to calculate the binary relation between each pixel and its local neighbor of a gray scale image. Eigenvalue of each pixel is calculated by extracting the binary number in clockwise direction and it is then converted into decimal value. Furthermore, Discrete Wavelet Transform (DWT) allows us to analyze images of different resolution. One of the advantages that makes DWT better than other similar transformation techniques such as Discrete Fourier Transform (DFT) and Discrete Cosine Transform (DCT) is that it contains both frequency and edge information. DCT is also one of the famous techniques for feature extraction. After DCT is performed on the image, three frequency components can be obtained, which are high, middle and low that contains some details and information in an image.

Edge detector algorithm -- Local Directional Patterns (LDP) is combined with the LBP as the modification for the feature extraction [4]. LBP is initially created for the texture description, but it is then enhanced for the face recognition due to high accuracy. For LBP, the image needs to be converted to grey scale, then each pixel will have a label that obtained by thresholding the 3 x 3 neighborhood of each pixel in line with the central pixel. Meanwhile for LDP, the image is divided into some smaller regions and LDP descriptors are obtained from the histograms. There are four classifiers used in this study. If only LBP is applied, the lowest accuracy of the recognition is around 93% for k-nearest neighbors (KNN) and the highest is 97.5% for AdaBoost classifier, while for LDP with Extend Local Binary Pattern (ELBP), the lowest accuracy obtained is around 94% for KNN and 99% for Voting Classifier. This study shows that the method of LDP combined with ELBP Feature Extractor with the Voting Classifier is able to improve the accuracy of facial recognition with LBP.

Local Binary Pattern Histogram (LBPH) and Haar Cascade classifier are also applied on low resolution images which from 76 x 76 pixels until 156 x 156 pixels in this study [5]. The histogram equalization and median filtering are used for the pre-processing of the face images. LBPH is an improved version of the LBP that use the histogram to represent the appearance of each binary code of the image. The proposed method successfully achieved a recognition rate of 99.67% for local files and 92.67% for real time images. As the recognition rate obtained is higher than the result from previous studies, this study proves that by using median filter and Haar Cascade classifier, the face recognition rate can be enhanced.

Viola-Jones algorithm is first used to detect the face in [6]. The authors apply LBP, Local Gradient Code (LGC), LDP, Histogram of Oriented Gradients (HOG) for image size of 256 x 256 with 16 x 16 blocks, 128 x 128 with 16 x 16 blocks and 128 x 128 with 8 x 8 blocks. The concept of LGC is according to the relationship of the neighboring pixels, while HOG is the illumination invariant magnitude or pixel orientation is applied to extract this information. The average recognition rate for LBP is 88.14%, for LGC is 88.8%, for HOG is 85.54% and for LDP is 77.69%. Overall, HOG, LBP and LGC have great performance in the feature extraction, but LDP performs the worst.

Viola-Jones algorithm and Haar Cascade features are used together for the face detection in [7]. LBPH is also involved for the conversion of the captured image into binary vector that enhance the face detection of Viola-Jones algorithm. Then, a pre-trained model of CNN, VGG16 is used as the classifier with the combination of max pooling and Softmax classifier. The reason of choosing CNN as the classifier is because it can avoid a clear functions distinction and indirect learning of training data. By using the dataset Karolinska Directed Emotional Faces (KDEF) as the validation, the model VGG-16 Face is able to achieve the accuracy of 88%. Viola-Jones algorithm has an invariant detector that locates scales, which means it allows the scaling the features instead of scaling the image itself. However, Viola-Jones algorithm is not effective to detect tilted faces and it is very sensitive to lightning conditions. While for Haar Cascade features, it consumes longer processing time than other face detection algorithm.

Principal component analysis (PCA) and LBP with the classifier support vector machines (SVM) and KNN with Euclidean distance (L2) are used in [8]. PCA and LBP both belong to one of the categories of feature extraction, which is also known as appearance-based techniques. Statistical approaches and linear transformation are used to process the whole face in order to search for the feature vectors that can illustrate the face. LBP divides the face image into several regions and AdaBoost is useful to select the most efficient LBP features. On the other hand, PCA involves in finding the orthogonal basis for data, arrange dimensions based on the significance and remove dimensions that are less important. Overall, PCA and LBP with SVM performs the best on JAFEE database and MUFE database, which the recognition rate is around 88% and 76%.

HOG and LBP are used for the feature extraction. Sparse Representation based Classification (SRC) is used due to its robustness to occlusions [9]. Since HOG extracts shape information primarily while LBP mainly extracts texture information, the authors proposed a method which is to combine HOG with SRC and LBP with SRC. For HOG, the image needs to be converted into grayscale image and the gradients is calculated. Next, it weighed vote into spatial and orientation cells. Contrast normalization is carried out over the spatial blocks and HOG is collected over the image now. Meanwhile for LBP extraction process, LBP code is calculated based on different LBP operator. Then it divides the image into local patches and each region will be used to construct a histogram. By combining the regional histogram to one histogram, the LBP feature is constructed. The result is calculated based on the Cohn-Kanade database. The True Positive (TP) of the confusion matrix for LBP with SRC and HOG with SRC are higher than 80%, while the fusion result of both methods achieves a higher accuracy which is 95.64%. However, in the proposed method, the suitable local patch size is difficult to be determined. Also, since SRC is used in this method, it is unsure that the other classifiers, for example the neural networks can have improvement on the result. Lastly, the time taken to process each image is also longer than those existing methods. The above methods are summarized in Table 1.

Table 1: Summary of emotion recognition by facial expression

Methods	Pros	Cons
CNN [1],	- Unique architecture from the constituent layer. - Simple network. - Avoid a clear function distinction and indirect learning of training data. - Flexibility to arrange/decide the layers.	- Slower training time due to extra operation such as max pooling. - Higher hardware requirement. - Require large dataset.
DBN [2]	- Able to look for model parameters quickly. - Variables can be computed efficiently in deepest layer.	- Ambiguous sensory input might not be able to interpret if top-down influences is ignored. - Only a layer of features is learnt at one time. - No parameters readjustment at lower-level layers. - Slow and inefficient. - Only consists of input and output layer.
LBP [3],[9]	- Transform image into an array based on the texture information. - Computational simplicity. - High discriminative power.	- Invariant to tilted images. - Increase of computational complexity in terms of time and space if the size features and neighbors increase. - Magnitude information is ignored.
DWT [3]	- Less information loss. - Both frequency and location information are captured.	- Depend on the number of decomposition levels. - Computational complexity and time. - Shift sensitivity. - Poor directionality. - Lack of phase information
DCT [3]	- Faster performance time. - Easier to implement and more efficient. - Able to store most information in fewer number of coefficients.	-
ELBP+ LDP+ Voting Classifier [4]	- Simple and efficient.	-
LBPH+ Haar Cascade [5]	- Better recognition rate than LBP. - Can be performed on low resolution images.	- The maximum number of faces that can be recognized in one frame is four only.
LDP [4],[6]	- Perform consistently if noise, illumination, expression and time lapse variations presented.	-
LGC [6]	- Obtained based on the relationship of neighboring pixels.	-
HOG [6]	- Illumination invariant.	- Sensitive to image rotation. - Only uses magnitude values of a pixel without considering neighboring pixels.
SVM [8]	- Perform faster prediction. - Use less memory.	- Long training time for large dataset. - Works poorly with overlapping classes. - Sensitive to type of kernel used.
KNN [4],[8]	-	- Slower performance for more features. - Does not work well with high dimensions. - Sensitive to noisy data, missing values and outliers.
Haar Cascade+ LBPH+CNN [7]	- Able to scale the features instead of scaling the image itself. - Avoid a clear function distinction and indirect learning of training data.	- Not effective in detecting turned faces. - Sensitive to lightning conditions. - Slower processing time.
PCA [8]	- Find the orthogonal basis for data. - Arrange dimensions based on the significance. - Remove dimensions that are less important.	-
HOG+SRC and LBP+SRC [9]	- Shape features and texture features are obtained. - Both features are fused in the end to achieve higher accuracy.	- Hard to determine the suitable size of the local regions. - It is unclear that the other classifiers can further improve the result.

B. Emotion recognition by voice

Linear Predictive Coding (LPC) is discussed in detail for the emotion recognition by speech or voice in [10]. It is a digital method to encode the analog signal which a linear function of the past values of the signal is used to predict a value. It is based on a mathematical approximation of the vocal tract. LPC consists of two main components: analysis (encoding) and synthesis (decoding). In the analysis part, LPC needs to first examine the speech signal. Then it will break it down into segments or blocks. Usually, a sender will carry out LPC analysis and transmit to receiver while the receiver performs LPC synthesis by building a filter that will reproduce the original speech signal. The strengths of this method are that it reduces the bitrates of the speech, requires less bandwidth thus increase the number of users and it uses the encryption of data. However, it has the drawbacks too. The quality of voice signal is reduced due to the reduction of the bitrates. This technique is a lossy compression technique. Hence, the data will fade away if it needs to transmit for a far distance.

Partila et al. [11] aim to detect stress in human speech by searching for a suitable combination of methods and features. Berlin database which contains ten sentences in the seven emotional states is used in their study and the dataset is passed into the Speech Emotion Recognition System (SERS), which includes the pre-processing, feature extraction and classification. After the emotion is classified, the recognition rate is investigated to find out which combination of feature extraction and classifier gives the best result. In the feature extraction, the parameters calculated including 39 Mel Frequency Cepstral Coefficient (MFCC), 12 LPC, 12 Linear Spectral Pairs (LSP) and 8 prosodic features (RMS energy, log-energy, zero crossing rate (ZCR), mean crossing rate (MCR), position of maximum, maximum, minimum, and harmonic-to-noise ratio (HNR). The first classifier used is Artificial Neural Network (ANN). As classifying emotional states can be considered as recognize patterns, it is possible to use a feed forward network of two layers with sigmoid hidden and output neurons. Scaled Conjugate Gradient (SCG) backpropagation is used to train this network. Next classifier involved is k-Nearest Neighbour which is a method of classification that uses the analogies learning principle. The classifier scans the space of training sample to determine the closest distance between known and unknown samples, then majority vote of the object will classify it into a category. Gaussian Mixture Model (GMM) is the last model used in the study. The Gaussian component densities' weighted sum represents the parametric probability density function. GMM is always applied in biometric system such system that recognizes speaker. For the experiment, MFCC features are used for the classification by these three models. Anger is recognized the most by the classifiers, sadness is recognized well too, but the state of fear and disgust have worst recognition rate by GMM and ANN. MFCC with ANN has the best recognition rate to recognize stress.

Kerkeni et al. [12] aim to compare the performance of SVM, Recurrent Neural Network (RNN) and Multivariate Linear Regression classification (MLR) on SER. They extract the MFCC and Modulation Spectral (MS) features from the speech signals. For some classifiers, they apply feature selection and speaker normalization to search for the relevant feature subset. Both regression and classification problems are solved by using MLR. The authors have modified the Linear Regression Classification (LRC) in MLR before passing the data to it for classification. In machine learning, SVM is an optimal margin classifier. If the experiment only has limited training data, SVM is suitable to be applied because it can better perform on the limited data. Lastly, RNN is excellent at learning time series data but at the same time, RNN faces the problem of gradient vanishing that increases when the training sequences increase. As the result, the highest accuracy of recognition is achieved by feature extraction of MFCC+MS with RNN, but without the speaker normalization, which is 94%.

Multilayer Perceptron (MLP), SVM and Logistic Regression (LR) with MFCCs are also compared in [13]. In the feature extraction process, three features are extracted, which are Mel-Spectrogram, chroma, and MFCC. Mel-Spectrogram contains Mel-Scaled frequencies and it takes emotion samples over time to act as audio signal. The signal is then mapped from domain of time and frequency using Fast Fourier Transform (FFT), and frequency and amplitude are shifted to construct the spectrogram. Chroma is used to extract the vocal content of the audio and it also determines the pitch rotation's angle as the helix traverses while MFCC is able to capture the phonetical crucial characteristics of audio file. As the result of the comparison, the accuracy of MLP model is 84.62%, LR is 100% while SVM is 91.67%.

Deep Neural Network (DNN) model is proposed to recognize emotional states from a one second frame of raw speech spectrogram because it contains acoustic features and semantic features. DNN is a feed forward neural network with many hidden layers and each layer consists of multiple neurons that hold a weight of the output of the

previous layer and an intercept term or bias. The result is passed to the next layer through a non-linear function such as sigmoid function, Softmax function and Rectified Linear Unit (ReLU). The weights during training are updated by using Mini-batch Stochastic Gradient Descent (SGD). Dropout is also included in this model since DNN is prone to overfit. The deep hierarchical architecture of DNN, augmentation of data and sensible regularization make the process of recognizing emotions from spectrogram achievable. Surrey Audio-Visual Expressed Emotion (SAVEE) database and eNTERFACE database are used to evaluate the accuracy of the proposed model. The accuracy of the emotion recognition by speech or voice is approximately 60% for both databases.

Harár et al. [14] proposes to use DNN with convolutional layers, pooling layers, dropout layers and fully-connected layers for Speech Emotion Recognition (SER). Angry, neutral and sad emotions of audio data in German Corpus (Berlin Database of Emotional Speech) are involved in their study. Voice Activity Detection (VAD) is applied so that the silent segments are eliminated, and SGD is used to optimize the DNN architecture. All audio data are standardized to have zero-unit variance and mean. The audio data are then split into training set, validation set and testing set according the ratio of 8:1:1. As for the result for the experiment, the DNN model has an accuracy of 96.97% on the testing data. It also achieved 69.55% average confidence on file prediction.

Another model of DNN, which is multi-task attention-based DNN model (MT-A-DNN) is proposed for SER [15]. The model is able to extract the multi-order dependency and sparseness evolved in the audio data, which makes it better than other plug-and-play models. The hierarchical multi-task framework of MT-A-DNN can study the latent structures from distinct feature perspectives since it shares the audio-data stream. A large-scale real-world database based on Chinese television shows and films is also established and used as the rich emotion materials. As the result, the model can predict the emotion correctly at an accuracy of 52.03% for anger, 67.73% for happiness, 60.18% for neutral and 60.25% for sadness. The average accuracy of MT-A-DNN is 60.022% that performs better than SVM, Random Forecast and DNN without attention mechanism and multi-task learning.

A deep Convolutional Neural Networks (CNN) from spectrograms for SER is proposed in [16]. Spectrogram is defined as a graphic depiction of the intensity of a signal at various frequencies from time to time in certain waveform. FFT is used to compute the speech signal in order to construct a time-frequency representation. Salient discriminative features are extracted from the spectrograms to carry out SER. One of the advantages of using spectrograms in the study is their robustness and discriminative features, which are automatically learnt from spectrograms as the basic for SER. The speech signals are generated to form the spectrograms as the inputs to CNN. For the CNN, there are three convolutional layers, several pooling layers, three fully-connected layers and finally softmax activation layer that carries out the classification task of seven emotions in the model. For the proposed model, it obtains the accuracy above 50% for anger, boredom, disgust and sad, while fear, happy and neutral have the accuracy below 50%. Compared to the pre-trained AlexNet model, only the emotions of anger, fear and neutral have better accuracy than the proposed model, while the other emotions achieved lower accuracy. Thus, the proposed model is concluded to have better performance and lesser complexity than the pre-trained model.

An Attentive CNN is proposed to compare the accuracy and efficiency of the model with traditional CNN for the emotion recognition by speech and voice in [17]. Attentive CNN is the combination of the strengths of CNN and attention mechanisms. Attentive CNN consists of a CNN with one convolutional layer to learn the representation of the audio signal, one pooling layer to avoid overfitting and an attention layer to compute the weighted sum of information extracted. Then, the output of the pooling layer and attention layer are passed to a fully-connected layer which has the softmax activation. They use different feature sets such as logMel filter-banks, MFCC, prosody feature set and the extended Geneva minimalistic acoustic parameter set (eGeMAPS) as the input into CNN and Attentive CNN. Overall, the Attentive CNN produces slightly higher accuracy then CNN which is around 62% on improvised sessions, 53.19% on scripted sessions and 55.5% on complete dataset. Both models work better with logMel, MFCC, and eGeMAPS, but slightly lower accuracy with prosodic features.

Other than using traditional classification methods on SER, a new method that combined CNN and RNN is proposed in [18]. In the conventional way, some low-level descriptors are extracted and they are used to train the machine. But it is very difficult for the researchers to select food features and optimize the mode. The authors mentioned that it is model-dependent to optimize the results for SER. However, the deep neural architecture is able to share those low-level structures and progress to high-level representations because of the network layers are stacked. Short Time Fourier Transform (STFT) is applied to generate the 2D representation from the speech signals after the pre-processing. The feature extraction of CNN with network layer of LSTM architecture will then analyze

the 2D representation. If only CNN is used for the SER, the accuracy of the model is 87.74%, while the accuracy of only LSTM (RNN) applied is lower, which is 79.87%. By joining the CNN with LSTM, which is also known as the time-distributed CNN, the accuracy is the highest, which is 88.01%.

A method of SER with CNN by using multi-task learning (MTL) is proposed in [19]. The authors mentioned that DNNs still have a generalization error problem because of limited training data, although DNNs are reported that it outperforms HMM and SVM. Hence, they propose MTL-based CNN (MTL-CNN) which is also called the transfer learning that classifies arousal level, valence level and gender as minor tasks and classifies emotion as the main task. They believe that with the help of those auxiliary tasks, it is possible to increase the performance of the model on the main task. For example, emotional states depend on the gender such as neutral women's voices might be similar and confused with highly aroused men's voices. Thus, emotion classification with gender specified can perform better than those recognize both genders. MTL-CNN consists of an input layer, two convolutional layers, one fully-connected layer and a MTL output layer. The accuracy of the emotion recognition by MTL-CNN with one main task and three auxiliary tasks is 89.59%, which is higher than the model that only focuses on the main task with the accuracy of 86.56%.

Zheng et al. [20] intend to carry out the emotion recognition by speech (or voice) on a CNN. They propose the method with the name PCA-CNNs-SER. For the pre-processing of the audio data, PCA is applied to decrease the dimensionality and suppress the interferences. CNN is constructed after non-overlapping segments are formed from the PCA whitened spectrogram. The proposed CNN is constructed by two convolutional layers, two pooling layers and two fully-connected layers. While Interactive Emotional Dyadic Motion Capture (IEMOCAP) database is used for the experiment, PCA-CNNs-SER achieves an accuracy of 40.02%, which outperforms the SVM classification with lower accuracy of 37.61%.

Two neural architectures, which are attention-based CNN-LSTM-DNN and CNN are implemented in [21]. First, MFCC is obtained, DCT is used to extract the MFCC and it is passed to the CNN. The convolutional layers extract the salient features while the sequential phenomena of the speech signal are handled by the Bi-directional Long Short-Term Memory (BLSTM) layers. A summary vector is extracted by the attention layer and passed as the input to the fully-connected dense layer (DNN) that has an output layer with the activation of softmax activation. The model is able to automatically and immediately learn the optimum representation of the voice signal from the raw time representation by applying LSTM network with CNN. The average of the overall accuracy of CNN-BLSTM-DNN is 87.2%, which is higher than CNN alone, which is 85%. However, the time consumption in training and classification in CNN-BLSTM-DNN is slower than CNN. The above methods are summarized in Table 2.

Table 2: Summary of emotion recognition by voice

Models	Pros	Cons
LPC [10],[11]	- Accurate estimation of the parameters of speech. - Efficient in computation of speech. - Reduce the bitrates of the speech. - Require less bandwidth thus increase the number of users. - Encryption of data.	- The quality of voice signal is reduced because of the reduction of the bitrates in voice signal. - Data faded away if it transmits too far.
MFCC [11],[12],[13],[17]	- Approximates the human system response more closely than other systems.	- For higher sample rates, the parameters are not able to reconstruct the speech samples back. - Inaccurate result if there is background noise.
ANN [11]	- Information stored in network, not in database. - Fault tolerance. - Output can be produced even with incomplete information	- Require processors with parallel processing power. - Network structure can only be built through experience, trial and error.

GMM [11]	- Learning speed is very fast.	- Estimating the covariance matrix becomes difficult when the data has insufficiently many points per mixture.
MLR [12]	- Simple and efficient. - Can be applied for classification and regression problems.	-
SVM [12],[13],[15], [19],[20]	- Good at classification especially for limited training data.	- Unable to produce probability output.
RNN [12],[18]	- Effective at learning temporal correlations - Good in learning time series data.	- Have vanishing gradient problem
MLP [13]	- Minimize risk of errors by adjusting parameters	-
DNN [15]	- Pipeline is relatively simpler.	- Requires large datasets. - Higher cost due to consumption of computing power. - Prone to overfit.
VAD+ DNN [14]	- Context independent.	- Unsure result for bigger dataset. - Only 3 classes of emotions can be classified.
MT-A-DNN [15]	- Learns the high-order dependency and non-linear correlations in audio efficiently.	- The database used for experiment is in Chinese, effectiveness of the model to other languages is unsure.
CNN with spectrograms [16]	- Learn robust and discriminative features from spectrograms automatically. - Flexibility to adjust. - Able to produce probability output.	-
CNN+LSTM (RNN) [18]	- Process is more complicated, but accuracy is only slightly better than CNN.	-
MTL-CNN (with three auxiliary tasks) [19]	- Reduce generalization error.	- Unrelated subtasks might decrease the performance of the model.
PCA-CNNs-SER [20]	- Reduce the dimensionality of the audio data.	- Suppress the interferences. - Produce low accuracy.
Attention-based CNN-BLSTM-DNN [21]	- The sequential phenomena of the speech signal are captured.	- Slower than CNN baseline.

III. Popularity of different models in emotion recognition through facial expression and voice

A comparison is made to compare the popularity of each classification model being used in emotion recognition through facial expression and voice. Note that the popularity is calculated on those models which are studied in this survey paper.

As shown in Figure 1, it is obvious that the most popular model is CNN, the deep learning model. It covered 25% of the research papers about emotion recognition through facial expression. CNN is a deep learning model which consists of feature extraction in it. By applying CNN, there is no need to apply the other feature extraction on the image dataset. CNN eases and smoothen the process of model training most of the time.

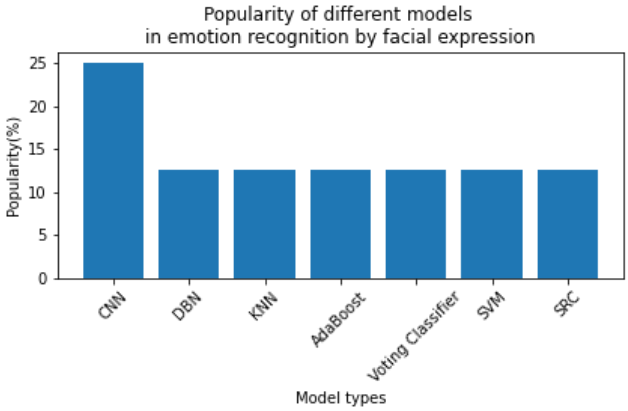

Figure 1. Popularity of different models in emotion recognition by facial expression

In Figure 2, the top model in all the research papers about emotion recognition through voice is also the CNN method. It gained a popularity of approximately 30%. It is believed that the popularity of CNN is due to its efficiency, high accuracy and flexibility to fine-tune. Many variations of CNN such as VGG-16 and combination with other methods are also applied. The second popular model is DNN with the popularity of around 20%.

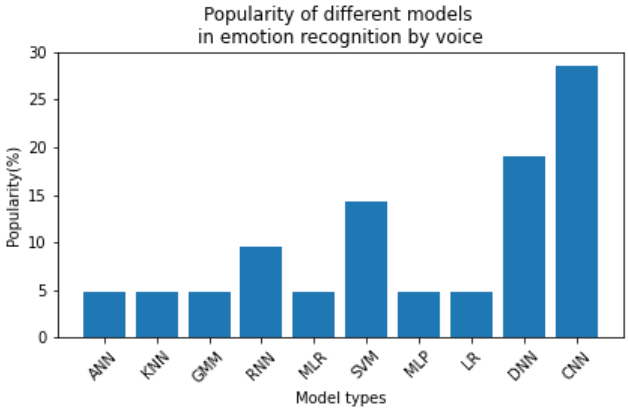

Figure 2. Popularity of different models in emotion recognition by voice

V. CONCLUSION

Among all of the classification models, CNN seems to be the most popular and widely used models when it comes to emotion recognition. It has the advantages of produce high accuracy result, feature extraction and flexibility to fine-tune. However, it also consumes more time and requires better resources in order to produce a good result. There are also other models that overcomes its drawbacks, but at the same time new problems can also arise. Different models can be selected and applied according to different circumstances. Hope this paper can cater the new researchers a quick identification on the most suitable method for recognizing the emotion through facial expression and/or voice.

ACKNOWLEDGEMENT

This work is supported by the funding of MMU Irfund (MMUI/220067).

REFERENCES

[1] Agrawal, A., & Mittal, ·. N. (2020). Using CNN for facial expression recognition: a study of the effects of kernel size and number of filters on accuracy. The Visual Computer 36 (2), 405-412. doi:10.1007/s00371-019-01630-9

[2] Agarwalla, N., Panda, D., & Modi, M. K. (2016). Deep Learning using Restricted Boltzmann Machines. International Journal of Computer Science & Information Security, 7(3), 1552-1556.

[3] Khan, S. A., Hussain, A., & Usmana, M. (2016). Facial expression recognition on real world face images using intelligent techniques: A survey. Optik, 127(15), 6195-6203. doi:10.1016/j.ijleo.2016.04.015

[4] Chengeta, K., & Viriri, S. (2018). A Survey on Facial Recognition based on local directional and local binary patterns. Conference on Information Communications Technology and Society (ICTAS). doi:10.1109/ICTAS.2018.8368757

[5] Isnanto, R. R., A. F., Eridani, D., & Cahyono, G. D. (2021). Multi-Object Face Recognition Using Local Binary Pattern Histogram and Haar Cascade Classifier on Low-Resolution Images. International Journal of Engineering and Technology Innovation, vol. 11, no. 1, 2021, 45-58. doi:10.46604/ijeti.2021.6174

[6] Kumaria, J., R.Rajesh, & KM.Pooja. (2015). Facial expression recognition: A survey. Procedia Computer Science 58, 486–491. doi:10.1016/j.procs.2015.08.011

[7] Hussain, S. A., & Balushi, A. S. (2020). A real time face emotion classification and recognition using deep learning model. Journal of Physics: Conference Series(Vol. 1432, No. 1, p. 012087). doi:10.1088/1742-6596/1432/1/012087

[8] Abdulrahman, M., & Eleyan, A. (2015). Facial Expression Recognition Using Support Vector Machines. 2015 23nd Signal Processing and Communications Applications Conference (SIU), 276-279. doi:10.1109/SIU.2015.7129813

[9] Ouyang, Y., Sang, N., & Huang, R. (2015). Accurate and robust facial expressions recognition by fusing multiple sparse representation based classifiers. Neurocomputing, 149, 71-78. doi:10.1016/j.neucom.2014.03.073

[10] Raja, M. N., Jangid, P. R., & Gulhane, S. M. (2015). Linear Predictive Coding. International Journal of Engineering Sciences & Research Technology.

[11] Partila, P., Voznak, M., & Tovarek, J. (2015). Pattern Recognition Methods and Features Selection for Speech Emotion Recognition System. The Scientific World Journal, 2015. doi:10.1155/2015/573068

[12] Kerkeni, L., Serrestou, Y., Mbarki, M., Raoof, K., Mahjoub, M. A., & Cleder, C. (2019). Automatic Speech Emotion Recognition Using Machine Learning. Social media and machine learning. IntechOpen. doi:10.5772/intechopen.84856

[13] Rumagit, R. Y., Alexander, G., & Saputra, I. F. (2021). Model Comparison in Speech Emotion Recognition for Indonesian Language. Procedia Computer Science, 179, 789-797. doi:https://doi.org/10.1016/j.procs.2021.01.098

[14] Harár, P., Burget, R., & Dutta, M. K. (2017). Speech Emotion Recognition with Deep Learning. 2017 4th International Conference on Signal Processing and Integrated Networks (SPIN), 137-140. doi:10.1109/SPIN.2017.8049931

[15] Ma, F., Gu, W., Zhang, W., Ni, S., Huang, S.-L., & Zhang, L. (2018). Speech Emotion Recognition via Attention-based DNN from Multi-Task Learning. Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems, 363-364. doi:10.1145/3274783.3275184

[16] Badshah, A. M., Ahmad, J., Rahim, N., & Baik, S. W. (2017). Speech Emotion Recognition from Spectrograms with Deep Convolutional Neural Network. 2017 international conference on platform technology and service (PlatCon), 1-5. doi:10.1109/PlatCon.2017.7883728

[17] Neumann, M., & Vu, N. T. (2017). Attentive Convolutional Neural Network based Speech Emotion Recognition:A Study on the Impact of Input Features, Signal Length, and Acted Speech. arXiv preprint arXiv:1706.00612.

[18] Lim, W., Jang, D., & Lee, T. (2016). Speech Emotion Recognition using Convolutional and Recurrent Neural Networks. 2016 Asia-Pacific signal and information processing association annual summit and conference (APSIPA), 1-4. doi:10.1109/APSIPA.2016.7820699

[19] Kim, N. K., Lee, J., Ha, H. K., Lee, G. W., Lee, J. H., & Kim, H. K. (2017). Speech emotion recognition based on multi-task learning using a convolutional neural network. 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), 704-707. doi:10.1109/APSIPA.2017.8282123

[20] Zheng, W. Q., Yu., J. S., & Zou, Y. X. (2015). An experimental study of speech emotion recognition based on deep convolutional neural networks. 2015 international conference on affective computing and intelligent interaction (ACII), 827-831. doi:10.1109/ACII.2015.7344669

[21] Hifny, Y., & Ali, A. (2019). Efficient Arabic emotion recognition using deep neural networks. ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 6710-6714. doi:10.1109/ICASSP.2019.8683632