SPEECH EMOTION RECOGNITION (SER) using Deep Neural Multilayer Perceptron (MLP)

Prashant Mishra
Sri Eshwar College of Engineering
Coimbatore, India

Nafees Ahmad
Sri Eshwar College of Engineering
Coimbatore, India

Naresh Kumar
Sri Eshwar College of Engineering
Coimbatore, India

Mr. A. Anandaraj, M.E. (Ph.D.)
Assistance professor , Sri Eshwar College of Engineering
Coimbatore, India

Abstract— **This paper presents an important and efficient method of Speech Emotion Recognition (SER) using Deep Neural Multilayer Perceptron (MLP). Speech Emotion Recognition has now become a necessity for human life. To achieve this study, many SER systems have been generated using various classifiers and different functions. Machine Learning has added efficient progress to this, by recognizing different speech graphs and similarities between them, and then Deep Learning is brought into consideration to study various graphs and perform feature extraction from the speech signals which is used for training different classifiers and retrieving data. Feature extraction was applied to get the most suitable dataset and use it for the most accurate results. The first recurrent neural network (RN) classifier is used to classify the seven emotions. Their performance is later compared with multivariate linear regression (MLR) and support vector machines (SVM) techniques, which are widely used in the field of emotion recognition for spoken audio signals. Berlin and Spanish databases are used as experimental data sets. This study shows that Berlin achieves 83% accuracy of all classifications for the database when a speaker normalization (SN) and a selection of features are applied to the features. For Spanish databases, optimal accuracy (% S%) is achieved by RNN classification without SN and by FS.**

I. INTRODUCTION

Emotions – They play an important role in daily interpersonal human interaction. This is necessary for our rational as well as intelligent decisions. It helps to match and understand the feelings of others by expressing our feelings and reacting to others. Research has shown that emotions play a powerful role in shaping human social interactions. Emotional performance conveys significant information about a person's mental state. This has opened up a new field of research called Automatic Spirit Recognition, with the original goal of understanding and achieving the desired emotions. In previous studies, many methods have been discovered to identify emotional states such as facial expressions, speech, physical gestures, etc. Some inherent advantages make speech cues a source for emotional computing. For example, compared to many other biological signals (e.g., electrocardiogram), speech signals can usually be obtained more easily and economically. That is why most researchers are interested in speech emotion recognition (SER).

The SER aims to identify the emotional state of the speaker from his voice. The area has received increasing interest in research in recent years. There are many applications to explore a person's emotions in interfaces such as robots, audio surveillance, web-based e-learning, commercial applications, clinical studies, entertainment, banking, call center. cardboard systems, computer games, etc. Orchestration or e-learning in the classroom, information about the emotional state of students can focus on increasing the quality of teaching. For example, the teacher may use SER to determine which subjects can be taught and should be able to develop strategies to manage emotions within the learning environment. That is why the emotional state of the learner in the classroom should be taken into consideration.

Deep neural networks (DNNs) are based on feed-forward structures composed of one or more underlying hidden layers between inputs and outputs. Feed-forward architectures such as Deep Neural Network (DNN) and Convolutional Neural Network (CNN) provide efficient results for image and video processing. On the other hand, repetitive architectures such as recurrent neural networks (RNN) and long short-term memory (LSTM) are very effective in speech-based classifications such as natural language processing (NLP) and SER [1]. In addition to their effective method of classification, these models have some limitations. For instance, the positive side of CNN is learning features from high-dimensional input data, but on the other hand, it also learns features from small changes and distortion events and, therefore, requires large storage capacity. Similarly, LSTM based RNN, variable input data, and model are capable of handling long-distance sequential text data.

There are three key issues that need to be addressed for a successful SER system, namely, (1) selecting a positive emotional speech database, (2) collecting effective features, and (3) creating reliable classifications using a machine learning algorithm. In fact, emotional feature extraction is a major problem in the SER system. Research Power, Pitch, Format Frequency, Linear Prediction Spectrum Coefficients (LPCC), Mel-Frequency Spectrum Coefficients (MFCC), Modulation Spectral Features (MSF) [5]]. Therefore, many researchers prefer to use the integrated feature set, which has a wide variety of features that contain more emotional information [6]. However, the use of an integrated feature set can lead to high quality and repetition of speech features. Therefore for most machine learning algorithms it complicates the learning process and increases the risk of over-fitting. Therefore, feature selection is inevitable to reduce the frequency of symptoms. Feature selection is presented in a review of models and methods [7]. Feature extraction and feature selection have the potential to improve learning performance, reduce computational complexity, create better generalization models, and reduce required storage. The final step in speech emotional recognition is classification. It involves classifying raw data into a specific type of emotion in the form of pronunciation or in the framework of pronunciation based on the characteristics extracted from the data. In recent years of speech emotional recognition, researchers have developed the Gaussian hybrid model (GMM) [8], the latent Markov model (HMM) [9], the support vector machine (SVM) [10, 11, 12, 13, 14], and neural networks (NN). [15], Repetitive Neural Networks (RNN) [16, 17, 18]. Some researchers have suggested other types of classifications, such as the modified Brain Emotional Learning Model (BEL) [19], which includes the Adaptive Neuro-Fuse Infection System (ANFIS) and Multilayer Perceptron (MLP) for speech-emotional recognition. Another specific strategy is the Multiple Kernel Process Process (GP) classification, which combines a linear kernel and a radial base function (RBF) kernel and provides two concepts similar to the learning algorithm. The Voice Segment Selection (VSS) algorithm proposed in [20] treats the voice signal segment as a different format image processing feature. It uses log-gabor filters to extract sound and unknown properties from the spectrogram for classification.

II. EXTRACTION

A. *Deep Learning*

Speech processing usually works directly on the audio signal. This is considered important and necessary for speech-based applications such as SER, speech denoting and music classification. With the recent developments, SER has gained a lot of importance. However, precise methods are still needed to mimic human-like behavior in communication with humans. As discussed earlier, the SER system includes various components such as feature selection and extraction, feature classification, acoustic modeling, unit recognition and especially language based modeling. Traditional SER systems typically have different taxonomic models, such as GMM and HMM. GMMs are used to capture the sound characteristics of sound units, whereas HMMs are used to handle transient changes in speech signals.

In-depth study methods involve various nonlinear factors calculated on a parallel basis. However, these techniques need to be adjusted using deeper layers of architecture to overcome the limitations of other technologies. Deep Boltzmann Machine (DBM), Repeated Neural Network (RNN),

Repeated Neural Network (RNN), Deep Believe Network (DBN), Convergence Neural Networks (CNLN) Basic depth exercises used to significantly improve SER.

B. *Recurrent Neural Networks*

The RNN is a segment of the neural network on which inputs and inputs are interdependent. In general, this interdependence is useful for estimating the future status of input. RNNs such as CNN require memory to store all the information obtained in the continuous process of deep learning modeling and only work effectively for certain back-propagation stages.

The main problem affecting the overall performance of the RNN is its sensitivity to the disappearance of gradients. In other words, the gradients in the training phase are significantly reduced and multiplied by a number of derivatives, small or large. However, this sensitivity decreases for a while and the inputs given at the initial stage are forgotten. To avoid such a situation, Long Short Term Memory (LSTM) is used to provide a block between repetitive connections. Each block of memory stores the temporary state of the network and contains gated units to control the flow of new information. The rest of the connections are usually very deep, so this can help reduce the gradient problem.

III. EMOTIONS RECOGNITIONS

Introduced in-depth learning method based on non-discriminatory pre-training method using DNN-HMM with MFCC modules. DNN-HMM is integrated with RNM using supervised training to detect different speech emotions. The hybrid in-depth learning method can achieve amazing results. Introduced the same DNN-HMM and compared it to the Gaussian hybrid model (GMM). This is investigated with a controlled Boltzmann machine (RBM) for supervised and indiscriminate pre-training. The results obtained in both cases were compared with the results obtained for both layers and with the multi-layer understanding of GMM-HMMs. Pre-trained hybrid DNN-HMMs received 12.22% supervised training, GMM-HMMs 11.67%, MLP-HMMs 10.56% and shallow NN2.2m accuracy. It refers to multimodality as an effective tool of research and includes a period to improve emotional recognition, strength and the efficiency of the recognition system.

The main problem affecting the overall performance of the RNN is its sensitivity to the disappearance of gradients. A custom SER system based on in-depth learning technology called DRN is used for SER. The learning phase of the model has frame-level and short-term acoustic characteristics due to the similar structure. Another multi-tasking deep neural network with hidden layers called MT-SHL-DNN that share feature exchange. Here, each data set output is supplemented with output layers. DNN helps to measure SER based on speaker behavior and gender. When DNs are used in nature to encode sections into vectors of a certain length, this is done by pooling different layers that are hidden at a given time. The design of the feature encoding process is used in conjunction with the segment level classification for effective classification.

The Convulsive Neural Network (CNN) also uses a layer-by-layer structure and can distinguish seven universal emotions from defined speech spectrograms. Introduced a technology for SER based on spectrograms and deep CNN. This model has three convoluted layers, which are fully integrated to extract emotional features from the spectrogram images of the speech signal. Another adaptation that implements the technology of sharing priorities between affiliate source and target classes (SPRST). A two-layered neural network that collects communication data from different sources and contexts often leads to inconveniences and thus destroys the overall performance of the system. Initially, pre-training of the weight is done for the first layer, and then the classification parameters of the second layer are implemented between the two classes taken. These classes with less data labeled in the target domain can retrieve information from the domain associated with the source to resolve errors.

Analyzes the tendency of DNs to learn specific features from different auditory emotional recognition systems. These features include voice and music based recognition. In addition, the use of cross-channel architecture improves overall

performance in a complex environment. This pattern gave some good results for human speech codes and musical codes; However, the results are not ideal for generalized auditory emotional recognition. The purpose of this cross-channel range is to capture specific features and connect them to a more generalized context. Additionally, these models can be combined with video-based DNNs to improve automatic SER. In such a case, the use of RNNs can further enhance time-based input data performance.

The study evaluates CNN using a more autonomous scenario called human-robot interaction (HRI). HRI used a humanoid robotic head, resulting in the necessary emotional feedback. It works with many processes at once: feature extraction based on information about shape, facial features and subject movement. Reception of internally interrupted fields increased the depth of the network for the extraction of emotions. Next, the cross-channel study is used to connect constant and dynamic currents in the same situation. As this model works effectively for the person performing automatic live expressions, it can be further extended to a multimodal system where visual stimuli can also be used as input with audio.

Hybrid deep learning modality can claim the inherent characteristics of RNN using CNN, allowing the model to achieve frequency and transient dependence on a given speech signal. In some cases, memory-enhanced reconstruction-error-based RNN can be used for continuous speech emotional recognition. This RNN model uses two components, the first for auto-encoder for feature reconstruction and the second for emotion prediction. It can also be used to gain a better understanding of the behavior of BLSTM-based RNN using regression models such as SVR.

IV. RESULT AND ANALYSIS

RNN is a segment of the neural network on which inputs and inputs are interdependent. In general, this interdependence is useful for estimating the future status of input. RNNs such as CNN require memory to store all the information obtained in the continuous process of deep learning modeling and only work effectively for certain back-propagation stages.

The main problem affecting the overall performance of the RNN is its sensitivity to the disappearance of gradients. In other words, the gradients in the training phase are significantly reduced and multiplied by a number of derivatives, small or large. However, this sensitivity decreases for a while and the inputs given at the initial stage are forgotten. To avoid such a situation, Long Short Term Memory (LSTM) is used to provide a block between repetitive connections. Each block of memory stores the temporary state of the network and contains gated units to control the flow of new information. The rest of the connections are usually very deep, so this can help reduce the gradient problem.

V. CONCLUSION

This paper provides a detailed overview of in-depth learning methods for SER. In-depth learning methods DBM, RNN, DBN, CNN and AE have been the subject of much research in recent years. These in-depth study methods and their layer-by-layer structure are briefly described based on the classification of various natural emotions such as happiness, joy, sadness, neutrality, surprise, boredom, hatred, fear and anger. These methods provide easy model training and shared load capability. The limitations of in-depth learning methods are their large layer-by-layer internal structure, the ability of temporarily changing input data, and the over-study of layer-by-layer information into memory or memory. This research work is a basis for assessing the performance and limitations of current in-depth learning methods. Additionally, it highlights some good directions for better SER systems.

It is still possible to strengthen the emotion-recognition system by integrating the database and classifier. The effectiveness of training multiple emotion detectors can be investigated by incorporating them into a single identification system. We aim for other feature selection methods as the quality of feature selection affects the rate of emotional recognition: a good emotion feature selection method can quickly select features that reflect the emotional state. The whole goal of our work is to develop a system that can be used for teaching interaction in the classroom, helping the

teacher orchestrate his or her class. To achieve this goal, we aim to test the system suggested in this work.

REFERENCES

[1] J. Deng, S. Frühholz, Z. Zhang and B. Schuller, "Recognizing emotions from whispered speech based on acoustic feature transfer learning", *IEEE Access*, vol. 5, pp. 5235-5246, 2017

[2] C. Huang, W. Gong, W. Fu and D. Feng, "A research of speech emotion recognition based on deep belief network and SVM", *Math. Problems Eng.*, vol. 2014, Aug. 2014

[3] M. S. Hossain and G. Muhammad, "Emotion recognition using deep learning approach from audio–visual emotional big data", *Inf. Fusion*, vol. 49, pp. 69-78, Sep. 2019.

[4] M. Chen, P. Zhou and G. Fortino, "Emotion communication system", *IEEE Access*, vol. 5, pp. 326-337, 2016.

[5] N. D. Lane and P. Georgiev, "Can deep learning revolutionize mobile sensing?", *Proc. ACM 16th Int. Workshop Mobile Comput. Syst. Appl.*, pp. 117-122, 2015.

[6] J. G. Rázuri, D. Sundgren, R. Rahmani, A. Moran, I. Bonet and A. Larsson, "Speech emotion recognition in emotional feedbackfor human-robot interaction", *Int. J. Adv. Res. Artif. Intell.*, vol. 4, no. 2, pp. 20-27, 2015.

[7] D. Le and E. M. Provost, "Emotion recognition from spontaneous speech using hidden MARKOV models with deep belief networks", *Proc. IEEE Workshop Autom. Speech Recognit. Understand.*, pp. 216-221, Dec. 2013.

[8] A. B. Nassif, I. Shahin, I. Attili, M. Azzeh and K. Shaalan, "Speech recognition using deep neural networks: A systematic review", *IEEE Access*, vol. 7, pp. 19143-19165, 2019.

[9] S. Lalitha, A. Madhavan, B. Bhushan and S. Saketh, "Speech emotion recognition", *Proc. Int. Conf. Adv. Electron. Comput. Commun. (ICAECC)*, pp. 1-4, Oct. 2014.

[10] K. R. Scherer, "What are emotions? And how can they be measured?", *Social Sci. Inf.*, vol. 44, no. 4, pp. 695-729, 2005.

[11] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis and S. Kollias, "Emotion analysis in man-machine interaction systems", *Proc. Int. Workshop Mach. Learn. Multimodal Interact.*, pp. 318-328, 2004.

[12] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias, W. Fellenz, et al., "Emotion recognition in human-computer interaction", *IEEE Signal Process. Mag.*, vol. 18, no. 1, pp. 32-80, Jan. 2001.

[13] O. Kwon, K. Chan, J. Hao and T. Lee, "Emotion recognition by speech signal", *Proc. EUROSPEECH*, pp. 125-128, 2003.

[14] R. W. Picard, "Affective computing", 1995.

[15] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: A review", *Int. J. speech Technol.*, vol. 15, no. 2, pp. 99-117, 2012.

[16] M. El Ayadi, M. S. Kamel and F. Karray, "Survey on speech emotion recognition: Features, classification schemes, and databases", *Pattern Recognit.*, vol. 44, no. 3, pp. 572-587, 2011.

[17] A. D. Dileep and C. C. Sekhar, "GMM-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines", *IEEE Trans. neural Netw. Learn. Syst.*, vol. 25, no. 8, pp. 1421-1432, Aug. 2014.

[18] L. Deng and D. Yu, "Deep learning: Methods and applications", *Found. Trends Signal Process.*, vol. 7, no. 3, pp. 197-387, Jun. 2014.

[19] J. Schmidhuber, "Deep learning in neural networks: An overview", *Neural Netw.*, vol. 61, pp. 85-117, Jan. 2015.

[20] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition", *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, pp. 474-477, Jul. 2005.

[21] C.-N. Anagnostopoulos, T. Iliou and I. Giannoukos, "Features and classifiers for emotion recognition from speech: A survey from 2000 to 2011", *Artif. Intell. Rev.*, vol. 43, no. 2, pp. 155-177, 2015.