Speech Emotion Recognition Using CNN and LSTM- A Review of Literature

Tanaya Bhagde[1], Saqlein shaikh[2], Aditi Upadhyay[3], Mahesh Sawant[4], Dr. Amit Gadekar[5]

[1,2,3,4] Student, Department of Computer Engineering, SITRC, Sandip Foundation, Nashik

[5] Professor, Department of Computer Engineering, SITRC, Sandip Foundation, Nashik

Abstract

Emotion recognition is always a very tough job, particularly if we are recognizing an emotion by using a speech signal. Many remarkable research works have been done on emotion recognition using speech signals. The primary challenges of emotion recognition are choosing the emotion recognition corpora (speech database), identification of different features related to speech, and an appropriate choice of a classification model. In this article, we use 13 MFCC (Mel Frequency Cepstral Coefficient) with 13 velocity and 13 acceleration components as features and a CNN (Convolution Neural Network) and LSTM (Long Short-Term Memory) based approach for classification. We chose Berlin Emotional Speech dataset (EmoDB) for classification purposes. We have approximately 80 percent of accuracy on test data.

KEYWORDS: CNN, Emotion recognition from speech, LSTM, MFCC

I. INTRODUCTION

As human beings' speech is amongst the most natural way to express ourselves. As emotions play a vital role in communication, the detection and analysis of the same area of vital importance in today's digital world of remote communication. Emotion detection is a challenging task because emotions are subjective. There is no common consensus on how to measure or categorize them. task. Classifying speech to emotion is challenging because of its subjective nature. This is easy to observe since this task can be challenging for humans, let alone machines. Potential applications for classifying speech to emotion are numerous, including but not exclusive to, call centres, AI assistants, counselling, and veracity tests.

In this project, we attempt to address these issues. We will use Convolutional Neural Networks and LSTM to classify opposing emotions. We separate the speech by speaker gender to investigate the relationship between gender and emotional content of speech. There are a variety of temporal and spectral features that can be extracted from human speech. We use statistics relating to the pitch, Mel Frequency Cepstral Coefficients (MFCCs), and Formants of speech as inputs to classification algorithms. The emotion recognition accuracy of these experiments allows us to explain which features carry the most emotional information and why.

II. LITERATURE REVIEW

SER system is primarily composed of three parts, i.e., the feature extraction, the feature selection, and the classifier. Some classifiers are mainly used in SER applications such as Support Vector Machine(SVM), Hidden Markov Model(HMM), Gaussian Mixture Model(GMM), Deep Neural Networks(DNNs) [6] [7] [8], etc. However, there has been no agreement on which classifier is the best choice for the SER system. So, comparing with the classifier, discriminative, robust, and affect-salient features' extraction and selection are much more important

Good features are the key to the study of SER, F. Wang et. al used the method of deep auto-encoder(DAE) that contained five hidden layers to extract features [9]. In contrast, they also extracted the traditional emotion speech features that included MFCC, Perceptual Linear Prediction cepstral coefficients(PLP), and LPCC from the speech signal. At last, the recognition results were obtained by using all the features as the input of the SVM model, the results showed that the speech emotion features extracted by the DAE had an obvious improvement over the others.

Z. T. Liu et.al proposed a framework of SER [10], firstly they builder the initial feature set which was composed of speaker-independent features and speaker-dependent features by extracting features. Secondly, selecting features by using correlation analysis that consists of distance analysis, partial correlation analysis, bivariate correlation analysis, and the Fisher criterion, then the redundant speech emotional features were discarded. After that, the optimal feature subset was obtained. Finally, an extreme learning machine(ELM) decision tree was constructed for emotion recognition. The experiment results showed that the ELM is more suitable for the decision tree algorithm and the effectiveness of the feature selection based on correlation analysis and the Fisher criterion was fully verified.

With the rapid development of deep learning, many researchers try to use deep learning to extract the speech emotion features. Because that the speech signal has the sequence characteristic, some researchers used temporal convolution networks(TCNs) and RNNs to process the speech signal. G. Trigeorgis et al. proposed the model that consists of two one-dimensional convolutional layers and two recurrent LSTM layers, which used the raw waveform signal as the input, the features obtained by the proposed method outperformed traditional designed features [11]. S. Mirsamadi et al. [12]proposed different RNN architectures for feature extraction, experiment results showed that using deep RNNs could learn temporal aggregation into longer time spans, furthermore a simple attention mechanism was added to the model, so the model could focus on emotionally salient parts of an utterance. Learning salient, discriminative features is an important research topic for SER, Q.R.Mao et al. [13]used spectrograms as the input of the model, the front part of the model was mainly about local invariant(LIF) learning that used CNN, and the remaining part used sparse auto-encoder(SAE) to analyze salient discriminative features, then affect-salient and discriminative features were obtained, finally the affect salient features were used to predict the emotion classes. A.M. Badshah et al. used a simple CNN model that consists of three convolutional layers and three fully connected layers to extract features from spectrograms, then predicted the emotion classes [14].

III. PURPOSE

Providing high-quality interaction between a human and a machine is a very challenging and active field of research with numerous applications. An important part of this domain is the recognition of human speech emotions by computer systems. In the past years, impressive progress has been achieved in speech recognition using deep learning. These achievements also include significant results on speech emotion recognition (SER). In this work, we build a neural network for SER on the IEMOCAP dataset and achieve the result highly competitive to the state of the art.

IV. PROPOSED ARCHITECTURE

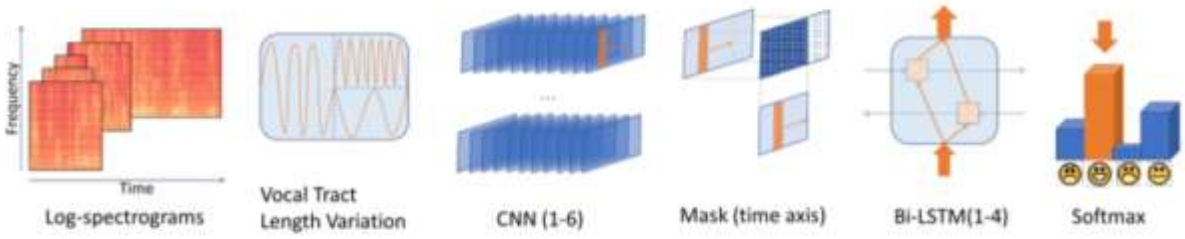

Figure 1 System Architecture

We referred to two datasets RAVDESS and SAVEE Dataset. Only took the audio data from this dataset. The RAVDESS database is gender-balanced consisting of 24professional actors. The speech part of the dataset includes calm, happy, sad, angry, fearful, surprise, and disgust expressions, and the song part of the dataset contains calm, happy, sad, angry, and fearful emotions. Each expression is produced at two levels of emotional intensity, with an additional neutral expression. We got 2000 audio samples which were in the .wav format. The SAVEE dataset consists of recordings from 4 male actors in 7 different emotions, 480 British English utterances in total. 'DC', 'JE',' JK', and 'KL' are the four male speakers recorded in the SAVEE database. There are 15 sentences for each of

the 7 emotion categories, and one file for each sentence. It includes 'anger', 'disgust', 'fear', 'happiness', 'neutral', 'sadness', and 'surprise'. The next step involves organizing the audio files. Each audio file has a unique identifier that tells the emotion of the file which can be used to determine the label of the audio file. We have 7 different emotions in our dataset. We used the Librosa library in Python to process and extract features from the audio files. Librosa is a python package for music and audio analysis. It provides the building blocks necessary to create music information retrieval systems. Using the Libros library we were able to extract features i.e MFCC(Mel Frequency Cepstral Coefficient).MFCCs are a feature widely used in automatic speech and speaker recognition. We also separated the female's and males' voices by using the identifiers provided on the website. This was because as an experiment we found out that separating male and female voices increased by 15 percentage. It could be because the pitch of the voice was affecting the results. Each audio file gave us many features which were an array of many values. These features were then appended by the labels which we created in the previous step.

V. CONCLUSION

Machine learning has made great progress so far, but in the field of speech signal processing, especially for building the SER system, there has not been much progress, SER is still a challenging problem. This paper proposes a new data enhancement method for SER and uses the proposed model that consists of the CNN, the LSTM, and the Attention Mechanism to classify speech emotions without using any traditional hand-crafted features. Comparing with the traditional method, this paper has achieved good results. In this paper, we just segment the .wav files into segments with the same length, this operation is not perfect. We are going to look for a better method to segment the .wav files more effectively and meaningfully

VI. REFERENCES

[1] Ashish B Ingale and DS Chaudhari. Speech emotion recognition. International Journal of Soft Computing and Engineering (IJSCE), 2(1):235–238, 2012.

[2] Moataz El Ayadi, Mohamed S. Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. Pattern Recognition, 44(3):572 – 587, 2011.

[3] Rainer Banse and Klaus R Scherer. Acoustic profiles in vocal emotion expression. Journal of personality and social psychology, 70(3):614, 1996.

[4] Christer Gobl and Ailbhe Nı Chasaide. The role of voice quality in communicating emotion, mood, and attitude. Speech communication, 40(1-2):189–212, 2003.

[5] Muhammad Huzaifah. Comparison of time-frequency representations for environmental sound classification using convolutional neural networks. arXiv preprint arXiv:1706.07156, 2017.

[6] Hamid Eghbal-Zadeh, Bernhard Lehner, Matthias Dorfer, and Gerhard Widmer. Cp-jku submissions for case-2016: A hybrid approach using binaural i-vectors and deep convolutional neural networks. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE), 2016.

[7] Erik Marchi, Dario Tonelli, Xinzhou Xu, Fabien Ringeval, Jun Deng, Stefano Squartini, and Bjorn W. Schuller. The up system for the 2016 ¨ case challenge using deep recurrent neural network and multiscale kernel subspace learning. 2016.

[8] Michele Valenti, Aleksandr Diment, Giambattista Parascandolo, Stefano Squartini, and Tuomas Virtanen. D case 2016 acoustic scene classification using convolutional neural networks. IEEE AASP Challenge on Detection and Classification of Acoustic Scenes and Events (DCASE2016), Budapest, Hungary, 2016.

[9] Wang Fei, Xiaofeng Ye, Zhaoyu Sun, Yujia Huang, Xing Zhang, and Shengxing Shang. Research on speech emotion recognition based on deep auto-encoder. In Cyber Technology in Automation, Control, and Intelligent Systems (CYBER), 2016 IEEE International Conference on, pages 308–312. IEEE, 2016.

[10] Zhen-Tao Liu, Min Wu, Wei-Hua Cao, Jun-Wei Mao, Jian-Ping Xu, and Guan-Zheng Tan. Speech emotion recognition based on feature selection and extreme learning machine decision tree. Neurocomputing, 273:271–280, 2018.

[11] George Trigeorgis, Fabien Ringeval, Raymond Brueckner, Erik Marchi, Mihalis A Nicolaou, Bjorn Schuller, and Stefanos Zafeiriou. Adieu ¨ features? end-to-end speech emotion recognition using a deep convolutional recurrent network. In Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on, pages 5200–5204. IEEE, 2016.

[12] Seyedmahdad Mirsamadi, Emad Barsoum, and Cha Zhang. Automatic speech emotion recognition using recurrent neural networks with local attention. In Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on, pages 2227–2231. IEEE, 2017.

[13] Qirong Mao, Ming Dong, Zhengwei Huang, and Yongzhao Zhan. Learning salient features for speech emotion recognition using convolutional neural networks. IEEE Transactions on Multimedia, 16(8):2203–2213, 2014.

[14] Abdul Malik Badshah, Jamil Ahmad, Nasir Rahim, and Sung Wook Baik. Speech emotion recognition from spectrograms with deep convolutional neural network. In Platform Technology and Service (PlatCon), 2017 International Conference on, pages 1–5. IEEE, 2017.

[15] Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory. ¨ Neural computation, 9(8):1735–1780, 1997.

[16] Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. Hierarchical attention networks for document classification. In Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 1480–1489, 2016.

[17] Yu Zhou, Yanqing Sun, Jianping Zhang, and Yonghong Yan. Speech emotion recognition using both spectral and prosodic features. In Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on, pages 1–4. IEEE, 2009.