SPEECH EMOTION RECOGNITION USING DEEP LEARNING

Vandana Singh
Department of Computer Science and Engineering,
Integral University, Lucknow.
K. C. Maurya
Assistant Professor Department of Computer Science and Engineering
Integral University, Lucknow.

Abstract

The purpose of this study is to detect the emotions evoked by the speaker while they are speaking. Speech generated in a condition of fear, rage, or delight, for example, becomes loud and quick, with a greater and broader range of pitch, but speech produced in a state of grief or exhaustion is sluggish and low-pitched. The detection of human emotions via voice and speech patterns has a variety of applications, including improving human-machine interactions. We provide a classification model of emotions produced by speeches that uses deep neural networks (CNNs), Support Vector Machine (SVM), and Multilayer Perceptron (MLP) Classification based on auditory data like Mel Frequency Campestral Coefficient (MFCC). . The models have been taught to distinguish between seven different emotions (neutral, calm, happy, sad, angry, fearful, disgust, surprise). Using the Ryerson Audio-Visual Dataset of Emotions Speech and Song (RAVDESS) dataset and the Toronto Emotional Speech Set (TESS) dataset, we found that the suggested technique achieves accuracies of 86 percent, 84 percent, and 82 percent using CNN, MLP, and SVM, respectively, for 7 emotions.

Index Terms-- Emotion detection, deep learning, machine learning, classification, mel-frequency cepstral coefficients, CNN, RAVDESS, TESS, SVM, MLP.

I. INTRODUCTION

The foundation for information exchange is human communication via spoken language. It is also used in a variety of practical applications in fields such as Business Process Outsourcing (BPO) Centers and Call Centers to detect emotion, which is useful for determining a customer's happiness with a product, improving speech interaction, resolving various language ambiguities, and adapting computer systems to an individual's mood and emotion.

The goal of the presented models is to identify just the emotion in the audio recording that has a higher value. To have a computer classify sentiments, several ways have been attempted, such as feature extraction or text analytics. The purpose of this study is use pure audio data while considering MFCC [4].

Detecting emotions is one of the most important marketing strategy in today's world. You could personalize different things for an individual specifically to suit their interest. For this reason, we decided to do a project where we could detect a person's emotions just by their voice which will let us manage many AI related applications. Some examples could be including call centers to play music when one is angry on the call. Another could be a smart car slowing down when one is angry or fearful. As a result this type of application has much potential in the world that would benefit companies and also even safety to consumers.

II. LITERATURE REVIEW

Many categorization algorithms have been proposed in this field of research throughout the years. Iqbal et al. [1] created a programme that employed Gradient Boosting, KNN, and SVM to work on granular partitioning in the RAVDESS data to find differences based on gender, with overall accuracy ranging from 40% to 80% depending

on the job. Male recordings alone, female recordings only, and mixed recordings datasets were constructed. SVM and KNN have 100% recognition for all anger and neutrality in the RAVDESS (male) dataset, while Gradient Boosting outperformed SVM and KNN in excitement and melancholy. SVM obtains 100% accuracy with the same fury as half of the guys in the RAVDESS (female) dataset.

With accuracy of 87 percent and 100 percent, KNN performed well in the areas of rage and neutrality. When compared to other categories of tourists, KNN performed worse in happiness and sadness. With rage and neutrality, SVM and KNN performed much better than Gradient Boosting among the combined male and female data rates. KNN's performance was extremely depressing in terms of both happiness and grief. The classifiers' average performance in the male dataset is better than in the female dataset without SVM. SVM is more accurate for aggregated data than gender data sets. [2]. Obtained 66.41 percent accuracy in audio data and 90 percent accuracy in blending audio and video data using another method. The scientists trained three alternative depth networks using already processed picture data, including faces and audio waveforms: one for image data only, one for fixed audio waveforms only, and one for both data and waveform data. One of the first algorithms to use the RAVDESS dataset, however it merely identified it from other emotions available [8]. Three different forms of music sharing algorithms have been proposed: a basic model, a single work area model, and a multi-task capacity model. A single, independently domain classifier was utilized in a basic model. During the training, two hierarchical kinds were employed. For each domain, the single function machine trained various classifications.

III. OBJECTIVE OF THE PROJECT

During the collecting step, noise typically corrupts the input data acquired for emotion recognition [4]. The extraction of features and categorization become less accurate as a result of these flaws [7]. This means that in emotions detection and identification systems, improving the data input is crucial. The emotional discrimination is retained in this pre - processing stage, but the speech and recording variance is removed [28].

The study will cover several deep learning algorithms in the context of SER in the next part. In comparison to traditional procedures, these methods produce more precise findings, but they are more computationally demanding. This section offers researchers and readers literature-based support for evaluating HCI(Human Computer Interaction) feasibility and analyzing the user's emotional voice in a specific scenario. Emotion identification from voice input data is a viable alternative [6], but real-time implementations of these approaches are far more challenging. Although these approaches have limitations, combining two or more of these classifiers creates a new step that may enhance emotion recognition.

IV. PROBLEM DEFINITION

On the RAVDESS and TESS datasets, the assessment findings reveal that the model is effective when compared to baseline methods and the state of the art. Table I. displays the accuracy, recall, and F1 values achieved for each of the emotional classifications. These findings demonstrate that accuracy and recall are extremely well matched, allowing us to acquire F1 values for practically all classes that are spread around the value 0.85. The model's robustness is demonstrated by the limited range of F1 values, which efficiently categories emotions into eight separate categories. The model is less accurate in the classes "Calm" and "Disgust," but this is not surprising given that they are the most difficult to recognize not just by speaking but also by monitoring facial expressions or analyzing written material [15], as stated in the Introduction. We chose to examine the findings acquired from two additional methods, namely SVM and MLP classifier, in order to assess the effectiveness of the emotion classification described in this paper.

V. PROPOSED METHODOLOGY

The emotion recognition classification models given here are based on a deep learning method based on CNN, SVM, and MLP classifiers. The fundamental concept is that the MFCC [4], often known as the "spectrum of a spectrum," is the only feature used to train the model. The Mel-frequency cepstrum (MFC) is a distinct understanding of the Mel-frequency cepstrum (MFCC), and it has been shown to be the state of the art in sound formalization in voice recognition [5]. Because of its capacity to express the amplitude spectrum of a sound wave in a condensed vectorial form, the MFC coefficients have been widely employed.

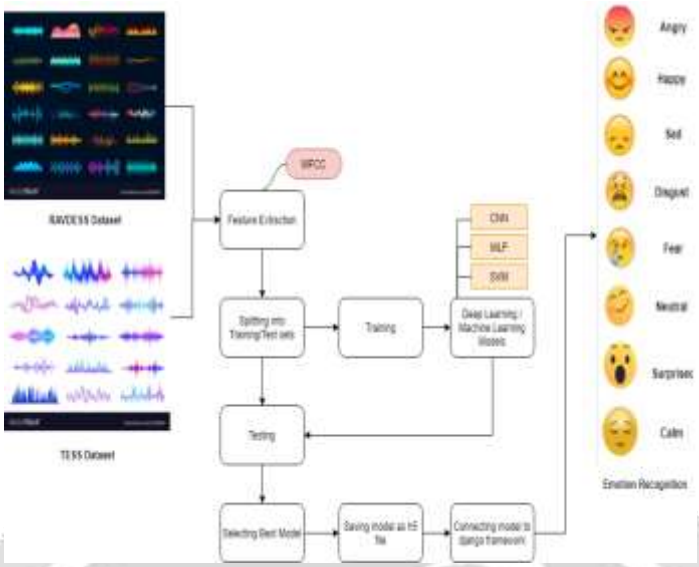

Fig-1: Proposed System Architecture.

Fig-2: Proposed CNN Layer Discription.

Method:

Figure 2 shows the operational results of the deep neural network (CNN) constructed for the classification job. For each audio file supplied as input, the network is capable of working on 40 feature vectors. The 40 values reflect the two-second audio frame's condensed numerical representation. As a result, we give a set of training data (40 x 1) on which we ran one cycle of a 1D CNN with a ReLU activation function [6, a 20% dropout, and a 2 x 2 max-pooling function. The rectified linear unit (ReLU) is defined as g(z) = max0, z, and it allows us to acquire a big value in the event of activation by using this function to represent hidden units. In this situation, pooling can assist the model in focusing solely on the most important aspects of each segment of input, rendering them position invariant. By adjusting the kernel size, we repeated the method outlined above. We then applied another washout and flattened the result to ensure compatibility with the next layers.

VI. REFERENCE

[1] Iqbal, A. and Barua, K. A real-time emotion recognition from speech using gradient boosting. In 2019 International Conference on Electrical, Computer and Communication Engineering (ECCE) (2019), IEEE, pp. 1–.

[2] Jannat, R., Tynes, I., Lime, L. L., Adorno, J., and Canavan, S. Ubiquitous emotion recognition using audio and video data. In Proceedings of the 2018 ACM International Joint Conference and 2018 International Symposium on Pervasive and Ubiquitous Computing and Wearable Computers (2018), ACM, pp. 956–959.

[3] LIVINGSTONE, S. R., AND RUSSO, F. A. : The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in north american english. PloS one 13, 5 (2018), e0196391.

[4] Logan, B., et al.: Mel frequency cepstral coefficients for music modeling. In ISMIR (2000), vol. 270, pp. 1–11.

[5] Muda, L., Begam, M., and Elamvazuthi, I. : Voice recognition algorithms using Mel Frequency Cepstral Coefficient (MFCC) and Dynamic Time Warping (DTW) techniques. arXiv preprint arXiv:1003.4083 (2010).

[6] Nair, V., and Hinton, G. E. : Rectified linear units improve restricted boltzmann machines. In Proceedings of the 27th international conference on machine learning (ICML-10) (2010), pp. 807–814

[7] Platt, J. C., Cristianini, N. and Shawe-Taylor, J. : Large margin dags for multiclass classification. In Advances in Neural Information Processing Systems 12, S. A. Solla,

[8] T. K. Leen, and K. Muller, Eds. MIT Press, 2000, pp. 547–553

[9] Toronto emotional speech set (TESS) (https://tspace.library.utoronto.ca/handle/1807/24487)

[10] S. Lalitha, A. Madhavan, B. Bhushan, and S. Saketh, "Speech emotion recognition," in Advances in Electronics, Computers and Communica- tions (ICAECC), 2014 International Conference on. IEEE, 2014, pp. 1–4.

[11] K. R. Scherer, "What are emotions? and how can they be measured?" Social science information, vol. 44, no. 4, pp. 695–729, 2005.

[12] T. Balomenos, A. Raouzaiou, S. Ioannou, A. Drosopoulos, K. Karpouzis, and S. Kollias, "Emotion analysis in man-machine interaction systems," in International Workshop on Machine Learning for Multimodal Interac- tion. Springer, 2004, pp. 318–328.

[13] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. Kollias,

[14] W. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," IEEE Signal processing magazine, vol. 18, no. 1, pp. 32–80, 2001.

[15] O.-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in Eighth European Conference on Speech Communica- tion and Technology, 2003.

[16] R. W. Picard, Affective computing. Perceptual Computing Section, Media Laboratory, Massachusetts Institute of Technology, 1995.

[17] S. G. Koolagudi and K. S. Rao, "Emotion recognition from speech: a review," International journal of speech technology, vol. 15, no. 2, pp. 99–117, 2012.

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

[19] .-W. Kwon, K. Chan, J. Hao, and T.-W. Lee, "Emotion recognition by speech signals," in Eighth European Conference on Speech Communica- tion and Technology, 2003.

[20] 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 transactions on neural networks and learning systems, vol. 25, no. 8, pp. 1421–1432, 2014.

[21] L. Deng, D. Yu et al., "Deep learning: methods and applications," Foundations and Trends® in Signal Processing, vol. 7, no. 3-4, pp. 197– 387, 2014.

[22] J. Schmidhuber, "Deep learning in neural networks: An overview," Neu- ral networks, vol. 61, pp. 85–117, 2015.

[23] T. Vogt and E. André, "Comparing feature sets for acted and spontaneous speech in view of automatic emotion recognition," in Multimedia and Expo, 2005. ICME 2005. IEEE International Conference on. IEEE, 2005, pp. 474–477.

[24] C.-N. Anagnostopoulos, T. Iliou, and I. Giannoukos, "Features and classifiers for emotion recognition from speech: a survey from 2000 to 2011," Artificial Intelligence Review, vol. 43, no. 2, pp. 155–177, 2015.

[25] [A. Batliner, B. Schuller, D. Seppi, S. Steidl, L. Devillers, L. Vidrascu,

[26] T. Vogt, V. Aharonson, and N. Amir, "The automatic recognition of emotions in speech," in Emotion-Oriented Systems. Springer, 2011, pp. 71–99.

[27] E. Mower, M. J. Mataric, and S. Narayanan, "A framework for automatic human emotion classification using emotion profiles," IEEE Transactions on Audio, Speech, and Language Processing, vol. 19, no. 5, pp. 1057– 1070, 2011.

[28] J. Han, Z. Zhang, F. Ringeval, and B. Schuller, "Prediction-based learning for continuous emotion

recognition in speech," in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017, pp. 5005–5009.

[29] "Reconstruction-error-based learning for continuous emotion recognition in speech," in Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on. IEEE, 2017.

[30] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," IEEE transactions on pattern analysis and machine intelligence, vol. 31, no. 1, pp. 39–58, 2009.

[31] T. Vogt, E. André, and J. Wagner, "Automatic recognition of emotions from speech: a review of the literature and recommendations for prac- tical realisation," in Affect and emotion in human-computer interaction. Springer, 2008, pp. 75–91.

[32] 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.

[33] S. Demircan and H. Kahramanlı, "Feature extraction from speech data for emotion recognition," Journal of advances in Computer Networks, vol. 2, no. 1, pp. 28–30, 2014.

[34] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in Fourth International Conference on Spoken Language Processing, 1996.

[35] Y. Zhou, Y. Sun, J. Zhang, and Y. Yan, "Speech emotion recognition using both spectral and prosodic features," in Information Engineering and Computer Science, 2009. ICIECS 2009. International Conference on. IEEE, 2009, pp. 1–4.