International Academic Institute for Science and Technology

International Academic Journal of Science and Engineering
Vol. 7, No. 2, 2020, pp. 18-22.

ISSN 2454-3896

International Academic Journal of Science and Engineering

www.iaiest.com

Design and Evaluation of Automatic Speech Recognition Model Using Si Base Dl Model

Devendra Joshi [a], Harpreet Kaur [b]

[a] *Devendra Joshi, Department of Computer Science and Engineering, Frontier Institute of Technology, Rajkot*
[b] *Harpreet Kaur, Department of Computer Science and Engineering, Frontier Institute of Technology, Rajkot*

Abstract

The signals of speech consist of sequences of sounds [1]. The transitions between these sounds serve as symbolic representations of the information being conveyed. The way these sounds are arranged follows the rules of language. Linguistics is the field that studies these rules and their application in human communication. Phonetics, on the other hand, focuses on the sounds of speech. We can represent speech through the content of a given message. An acoustic waveform is another way to characterize speech, capturing the signal that carries its message. For humans and their environment, speech is vital for communication, which is why Automatic Speech Recognition (ASR) systems are essential. There are various types of speech, including isolated words, continuous speech, and connected speech. The study proposed for SER uses two supervised machine learning algorithms, support vector machines and decision trees, enhanced with swarm intelligence optimization techniques, such as PSO and WOA, and a deep learning methodology dubbed a deep belief network. The DBN automatically extracts features from the input speech signal after pre-processing. Given that voice signals frequently lose a lot of information, making performance questionable, this approach offers more dependable features.

Keywords: ASR, SI, DL, ML.

Received: 12 September 2020/Revised: 24 October 2020/Accepted: 06 November 2020/Published: 31 December 2020

Introduction

It's interesting to note that three hybrids displayed desirable substantial standard heterosis, whereas one hybrid was determined to have a negative significance for mid-parent and better parent heterosis [1]. In comparison to their mid-parent, two hybrids displayed a significant and negative heterosis in terms of plant height [2]. Four hybrids also showed significant negative better parent heterosis. Unfortunately, none of the hybrids displayed the kind of negative heterosis for plant height that we would consider desirable when looking at normal heterosis. In contrast, a significant positive mid-parent heterosis for the number of tillers per plant was seen in five hybrids. Five hybrids also displayed the desired positive heterosis when compared to their superior parent. Regarding normal heterosis, there was just one hybrid that stood out as having significant advantages. Moving on to the panicles per plant, eight hybrids exhibited a considerable positive mid-parent heterosis, while four hybrids showed a positive and substantial better parent heterosis. However, when it came to the quantity of generating panicles per plant, none of the hybrids performed better than the normal heterosis. For the 100-grain weight, twenty-four hybrids displayed a desirable positive substantial mid-parent heterosis [3]. On the other hand, fifteen hybrids exhibited noteworthy improvements in parent heterosis. Finally, 23 hybrids were determined to be desirable in terms of standard heterosis.

Eleven hybrids exhibited a considerable and positive mid-parent heterosis in terms of grain length. Four hybrids stood out as attractive when considering superior parent heterosis. Five hybrids were discovered to be favourable for conventional heterosis. In terms of grain breadth, 18 hybrids shown a positive and significant mid-parent heterosis. Similarly, 18 hybrids also demonstrated notable positive results for enhanced parent heterosis. Standard heterosis was desired in twenty-four hybrids. Nine hybrids were determined to be desirable for mid-parent heterosis based on leaf width, whereas seven hybrids were favourable and sought better parent heterosis. It was found that sixteen hybrids were suitable for traditional heterosis. Thirteen hybrids showed a positive significant mid-parent heterosis in terms of panicle length, while twenty-one hybrids showed a significant negative mid-parent heterosis [4]. While 25 hybrids showed significant negative heterosis, nine hybrids showed positive significant results for superior parent heterosis. It was thought desirable to have 14 hybrids with standard heterosis. Ultimately, for panicle effort, sixteen hybrids displayed a significant positive mid-parent heterosis, whereas twenty-two hybrids displayed a significant negative mid-parent heterosis. Thirty-two hybrids showed significant negative heterosis, while six hybrids showed significant positive findings for superior parent heterosis. According to standard heterosis, 19 hybrids were deemed suitable [5].

Literature Review

Using the Berlin EmoDB and IEMOCAP databases, researchers have identified a range of emotions, including happiness, sadness, neutrality, surprise, disgust, fear, and aggression. The results were impressive, with the Deep 1D and 2D CNN-LSTM models achieving accuracies of 81.6% and 82.9%, respectively. In another study, the IEMOCAP database was used to pinpoint emotions such as anger, joy, sorrow, and neutrality, employing an LSTM-based three-layer recurrent neural network that yielded an overall accuracy of 71.04%. High-level characteristics can be extracted from input audio and log-mel spectrograms using a combination of deep 1D and 2D CNNs [6]. LSTM might be essential in the future to increase SER accuracy.

In a brief timeframe, typically between 10ms to 30ms, we can apply spectral analysis methods to speech signals to capture the short-term spectral characteristics of the signal. There are several techniques available for spectral analysis, including Fast Fourier Transform (FFT), linear prediction, Discrete Fourier Transform (DFT), and filter banks. Notably, Atal et al. highlighted that spectral analysis is one of the most effective feature extraction methods, yielding impressive results. For example, the Cambridge University-developed HTK recognizer used signal energy and 12 Mel Frequency Cepstral Coefficients (MFCC) to generate a basic 13-element acoustic feature vector. Additionally, they demonstrated how MFCCs can greatly improve the functionality of the majority of voice recognition systems.

Materials and Methods

Since each utterance has unique characteristics in terms of energy and spectral features, it's essential to identify one or more utterances that best represent all members of the same class. In a speech recognition system, this procedure is called labeling and is commonly referred to as training or learning. Applying our previous example of identifying the Chinese word "one," we must train and record this particular utterance in the database [7]. For training purposes, the word "one" in this context can be represented by four sets of feature vectors that are similar yet different.

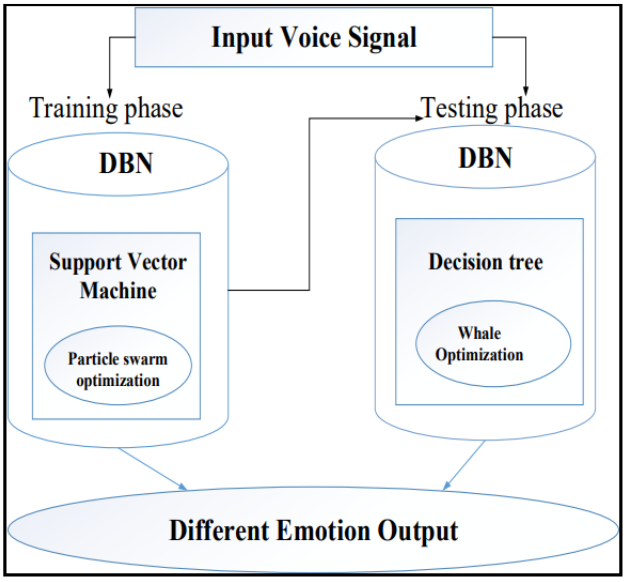

Fig. 1: Schematic Diagram

Dynamic interactions may provide a variety of emotion-specific experiences that stem from the same basic emotional state but have different perceptual preferences. In the subsequent emotion classification phase, the SVM classifier is employed for dual classification. To improve this classification process, we need to eliminate errors and achieve the best discrimination of the extracted attributes. For this reason, during the training phase, we combined the PSO technique with the SVM classifier. During the testing phase, a decision tree is used to accurately detect emotions, such as anger, disgust, fear, happiness, neutrality, and sorrow, in order to minimize the overfitting issue and the risks of relying just on one classifier [8].

Results and Discussions

The input preprocessing is complete, and features are extracted at the same time using the suggested SDCN. After that, these feature values are passed into an SVM classifier for classification. However, this method has its drawbacks, which are addressed through the PSO optimization approach. The goal of this combination is to save time. We compare this to a method that combines SVM-based classifiers with PSO optimization techniques to evaluate how effective the SVM classifier strategy really is. Additionally, the whale optimization technique helps simplify the complexities of the output generated by the decision tree.

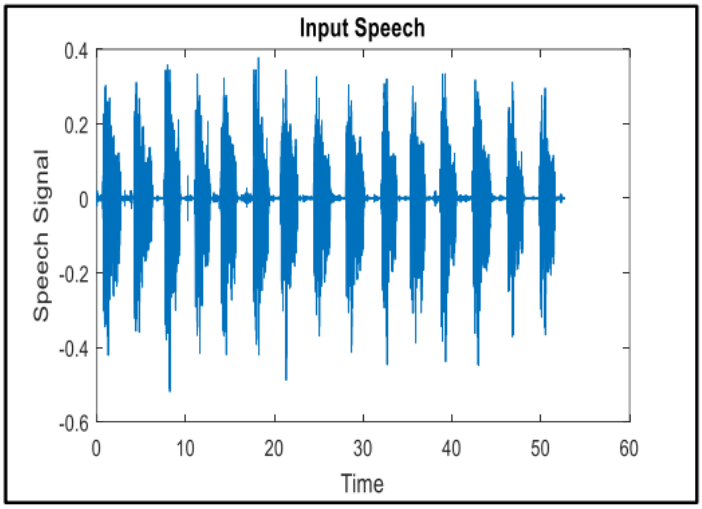

Fig. 2: Input Speech Database Waveform

The experimental findings demonstrate that the suggested set of algorithms improves speech verification and expedites the process of identifying sentiment in voice signals. Additionally, it has been demonstrated that this configuration converges faster than conventional gradient descent techniques.

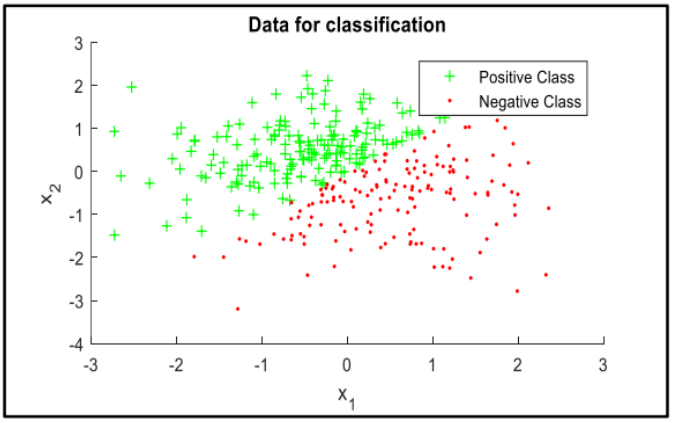

Fig 3: SVM Classifier and PSO Optimization Algorithm to Classify Positive and Negative Classes.

For a dialogue to be considered effective, the emotion must be accurately interpreted and responded to in a timely manner; however, the complexity of the selection tool can sometimes cause delays. An SVM classifier is employed to sort the data illustrated in Figure 3, using a specific threshold to organize everything neatly. This leads to the classification being represented in two distinct colors. When the values exceed the threshold, they pop up in green above a hyperplane. Conversely, if they fall below the threshold, they appear in red beneath the hyperplane, clearly showing the classification. The features for categorization are defined by X1 and X2. A 2D hyperplane is used to effectively demonstrate the classification, making it easy to visualize even when the data hovers around the threshold value.

Conclusion

In the world of speech recognition, we need to turn spoken words into digital formats so that computers can process them. This is done using devices like microphones, which capture sound waves and convert them into a series of digital data points. In order to preserve the integrity of the speech, it is imperative that this conversion be exact enough to replicate the original sound. Shannon's sampling theorem states that we won't lose any of the signal's integrity if we sample at the Nyquist rate, which is twice the original signal's bandwidth. A sampling rate of 8KHz is more than sufficient to recreate the original sound without any loss, as human speech normally occurs at frequencies lower than 4KHz. An analog speech stream can be converted into a digital format using an analog-to-digital converter with a sampling frequency of 16,000 Hz and a resolution of 12 bits.

References

Kim, M., Cao, B., Mau, T., & Wang, J. (2017). Speaker-independent silent speech recognition from flesh-point articulatory movements using an LSTM neural network. *IEEE/ACM transactions on audio, speech, and language processing*, *25*(12), 2323-2336. https://doi.org/10.1109/TASLP.2017.2758999

Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, *64*(4), 532-556. https://doi.org/10.1109/PROC.1976.10159

Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, *56*, 85-100. https://doi.org/10.1016/j.specom.2013.07.008

Lee, K. F., Hon, H. W., & Reddy, R. (1990). An overview of the SPHINX speech recognition system. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *38*(1), 35-45. https://doi.org/10.1109/29.45616

Yu, D., & Li, J. (2017). Recent progresses in deep learning based acoustic models. *IEEE/CAA Journal of automatica sinica*, *4*(3), 396-409. https://doi.org/10.1109/JAS.2017.7510508

Reddy, D. R. (1976). Speech recognition by machine: A review. Proceedings of the IEEE, 64(4), 501-531. https://doi.org/10.1109/PROC.1976.10158

Van Doremalen, J., Boves, L., Colpaert, J., Cucchiarini, C., & Strik, H. (2016). Evaluating automatic speech recognition-based language learning systems: A case study. *Computer Assisted Language Learning*, *29*(4), 833-851. https://doi.org/10.1109/JAS.2017.7510508

Karpagavalli, S., & Chandra, E. (2016). A review on automatic speech recognition architecture and approaches. *International Journal of Signal Processing, Image Processing and Pattern Recognition*, *9*(4), 393-404.