

Automatic Emotion Recognition Model Using DBN Model

Tarun Bansal ^a, Simran Arora ^b

^a *Department of Electrical Engineering, Royal Institute of Electronics, Gwalior, India.*

^b *Department of Electrical Engineering, Royal Institute of Electronics, Gwalior, India.*

Abstract

Machine learning techniques can now identify human emotions from voice signals, but their poorer detection accuracies in practical applications stem from their lack of rich representation capability. Speech signals include numerous levels of representations that can be automatically discovered by deep belief networks (DBN). This research proposes an ensemble of random deep belief networks (RDBN) technique for speech emotion recognition in order to fully utilize its benefits. After first extracting the input voice signal's low-level properties, it uses them to create a large number of random subspaces. After that, each random subspace is given to DBN, which uses the higher-level features as the classifier's input to produce an emotion label. The final emotion label for the input speech signal is then determined by fusing all of the outputted emotion labels using majority voting. Experimental findings on benchmark speech emotion datasets demonstrate that RDBN outperforms the comparison techniques in terms of speech emotion recognition accuracy.

Keywords: ASR, SI, DL, ML.

Introduction

In the fields of human-computer interaction (HCI), machines must be able to recognize human emotions in order to make decisions, give suitable feedback, and engage in personalized interactions with humans. Humans are able to communicate their feelings both orally and nonverbally. When interacting with humans, a machine must be able to distinguish between the two sorts of human emotions; the latter is referred to as human affective behavior (Besacier et al., 2014). In HCI, it is essential to identify emotional cues from their affective behaviors. Certain types of ideas and affective behaviors cause physical changes in the human brain. In order to comprehend ideas, researchers have been exploiting signals taken from the human brain that correlate with changes in physical processes (Jelinek, 1976). One common method for comprehending the physiological functions of the brain that makes use of electroencephalogram (EEG) signals is brain typing. Humans experience emotions at all times and in all places. They can be identified and expressed by speech cues, which make up 38% of all emotions expressed. Speech emotion recognition (SER), which automatically categorizes a speaker's emotional state from speech signals into one of several fundamental emotions, has lately gained attention for this reason. SER has been used to address problems in a variety of fields (Karpagavalli & Chandra, 2016).

This study employed multimodal physiological signals to identify human emotions in order to extract genuine feelings and get around the drawbacks of functional neuroimaging-based emotion detection techniques. The physiological data from Zygomaticus Electromyography (zEMG), Electro-Dermal Activity (EDA), and PPG sensors are combined to distinguish between five basic emotion classes: neutral, happy, relaxed, disgusted, and sad. We employed a deep belief network (DBN) to extract depth level features from the combined physiological outputs of three biosensors (Kim et al., 2017).

Literature Review

Furthermore, in order to achieve higher accuracy, neural network-based classification models need training data; yet, these models frequently incorporate low-level features, local ordering, and intrinsic characteristics that are difficult to control in auditory models (Lee et al., 1990). All things considered, the method should be created by (i) picking the most crucial signal characteristics to incorporate comprehensive information for any classification model to recognize effectively, and (ii) picking samples for the classification model's training with care. In recent studies, deep neural networks (DNN) have made impressive strides in both speech and image processing, although research on speech emotion recognition (SER) is still somewhat limited. It turns out that DNN offers significant advantages in recognizing human emotions, even if the results can sometimes be a bit off. In order to address this, we created a five-layer automated framework that can reliably identify and train emotions. In order to gain a better understanding of emotions, this paper presents a classifier framework. Regretfully, no theoretical framework exists that can definitively pinpoint the characteristics associated with the emotions expressed in speech signals. As a result, this approach focuses on finding ways to overcome the existing challenges in this area (Reddy, 1976).

Materials and Methods

These days, our interactions with technology have become quite conversational. The SER hardware is designed to pick up on our emotions and respond just like a human friend would. In order to improve the accuracy of emotion identification, we investigate a classification method in this work that focuses on a combination of four important emotions: anger, fear, sorrow, and disgust, or, in essence, stress, anxiety, and depression. Accurately identifying these feelings can be challenging. Therefore, motivated by the idea of Elaeocarpus Ganitrus, we present a novel deep ganitrus algorithm (DGA) that recognizes human emotions and sentiments from speech. The fruits of the Elaeocarpus Ganitrus tree contain this specific spherical jewel of a bead. The quantity of "mukhis"—the clefts and furrows on the surface of these Rudraksha beads—determines their classification. The most widely used beads are those with 1 to 14 mukhis, however older literature mention beads with 1 to 38 mukhis. The panche, or five-faceted, Mukhi Rudraksha bead is especially well-liked. Higher mukhi beads are rather uncommon, and depending on how many mukhis a bead contains, each one has a different effect. These beads are used in traditional medicine to treat psychosomatic diseases, heart palpitations, stress, anxiety, depression, and nerve pain (Van Doremalen et al., 2016). They are supposed to encourage relaxation and reduce body warmth. The fruit stone, or seed kernel, is not only delicious but also calming and pleasant. While the beads may look similar at first glance, a deeper classification based on their lines reveals a variety of types. Ultimately, these beads are considered a powerful medicinal resource. Just like humans, we experience a wide range of emotions. However, recognizing these emotions with SER is no easy feat. For instance, a voice signal might seem to convey a single emotion, but a closer look can uncover a mix of feelings. As illustrated in Figure 1, our proposed DGA method uses multiple techniques to differentiate between these emotional nuances.

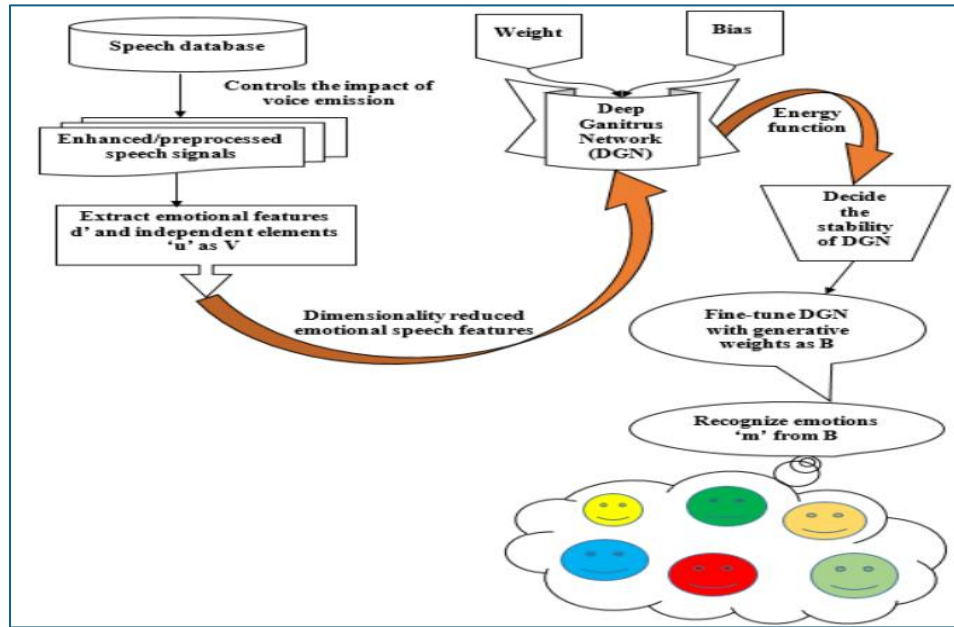


Fig. 1: Schematic Diagram

Two popular methods for analyzing big data sets are principal component analysis (PCA) and factor analysis (FA). Independent component analysis (ICA) goes one step further by identifying a set of genuinely independent source signals, whereas PCA and FA seek to detect signals that are less independent. PCA and FA specifically look for uncorrelated signals. Two fundamental presumptions underlie ICA's operations. First, for the hidden components to be successfully identified, they must be non-Gaussian and statistically independent. This independence in linguistics refers to the fact that understanding variable x does not provide insight into variable y , and vice versa. Second, the capacity to extract emotional traits as independent variables is a unique property of ICA. ICA places a higher priority on independence than PCA, which seeks to maximize the variance of data points. In essence, ICA is a method for splitting a multivariate input into additive subcomponents. This is accomplished by presuming that each component is a statistically independent non-Gaussian signal. Conversely, generative probabilistic models called Deep Belief Networks (DBN) have several layers of hidden variables, each of which captures important high-order correlations between the hidden characteristics in the layer underneath it (Yu & Li, 2017). These models have been effectively used in a variety of sectors because of effective approximate inference and greedy learning techniques. A DBN's fundamental component is the Restricted Boltzmann Machine (RBM), a bipartite undirected graphical model that has traits in common with each of the DBN's distinct layers. The next task after building the model is training it. The maximum-likelihood approach is frequently used to train the parameters of this probabilistic model. Unfortunately, the partition function, a computationally difficult normalizing constant, is the only way to determine the probability of the data presented by the model.

Results and Discussions

This section presents the suggested deep ganitrus algorithm, which is intended to extract discrete emotions. By isolating emotional speech parts, our technique heals people's speaking abilities and their cumulative capacities independently. After determining the final emotional element of the consultant's speech, it examines the passionate data that was gathered and eliminates any extraneous details. We also examine the results of comparing our implementation of emotion detection with other system-based methods. For example, a vocal signal that aids in distinguishing between various expressions is shown in Figure 2. Our goal is to develop an efficient, real-time framework for autonomous emotion recognition systems that will help consumers, call center workers, drivers, pilots, and anybody else interacting with a machine. In order to make robots appear and behave more like humans, we've discovered that giving them emotions is crucial.

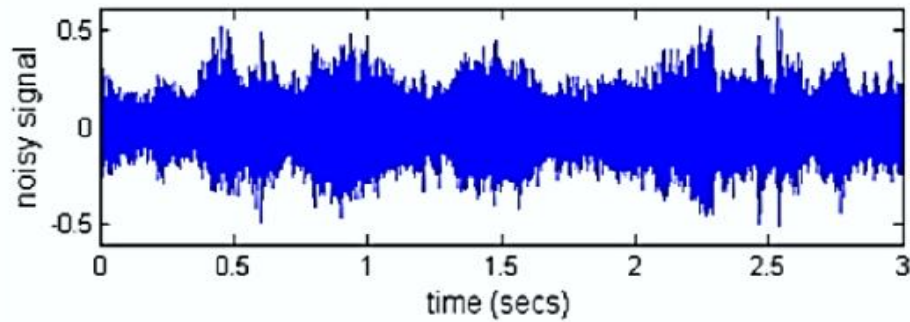


Fig. 2: Input Speech Database Waveform

Since speech transmits richer information, it is more effective to decipher a speaker's emotions from their speech signals. We can address the main issues with emotion identification systems by combining several spatial transformation techniques. For example, a variety of emotions, including anger, joy, sadness, fear, contempt, boredom, and neutrality, are frequently included into people's words when they speak. Figure 3 shows different ratios of the emotions extracted from a single speech input. The feelings depicted in Figure 3 were effectively identified in the sentence, "Currently at weekends I always went home and saw Agnes."

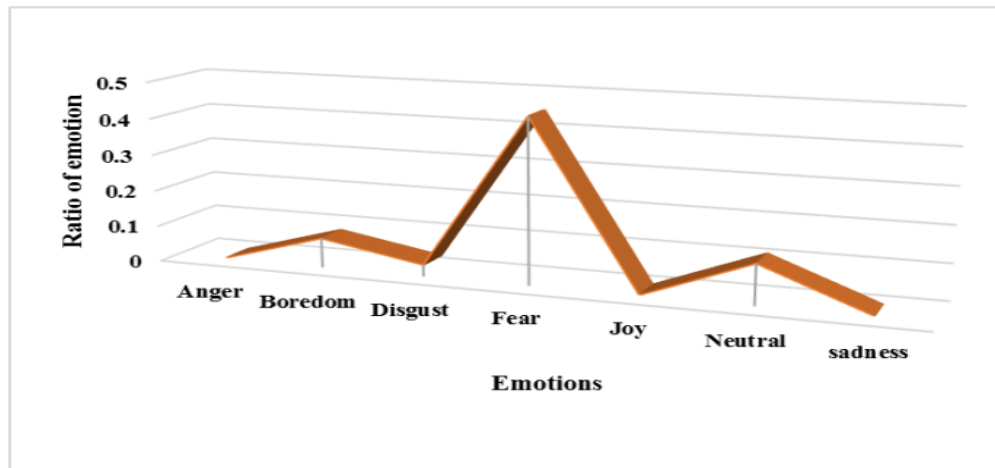


Fig. 3: Emotions Recognized from a Single Speech Signal

We tested the suggested DGA architecture on the Berlin emotional dataset. The Berlin database's speech emotions were recorded in a variety of speech variations and settings, which may have included some undesired noise. Our method is intended to automatically eliminate those bothersome messages.

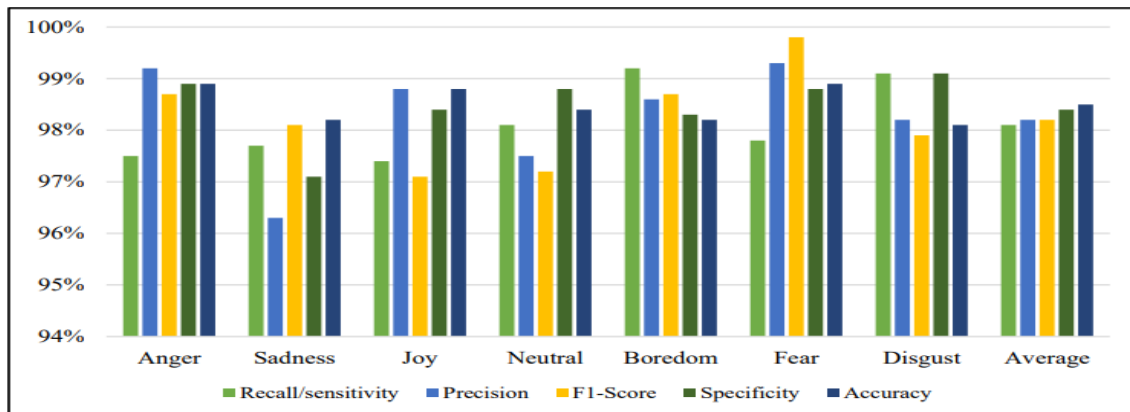


Fig. 4: Performance Metrics

We then applied the feature extraction process to the enhanced speech stream we had prepped earlier. After that, we used the Deep Ganitrus Algorithm to recognize emotions based on the features we recovered. In the end, it identifies the seven emotions found in the Berlin database.

Conclusion

Recognizing human emotions through speech signals has become a hot topic in recent years. It's quite a challenge, though, because factors like the speaker's style, voice quality, cultural background, and environment all play a role. Plus, emotions themselves are tricky to define, which complicates the task of identifying them through speech. Although numerous signal processing frameworks and techniques have been created to identify and evaluate emotions in speech, they frequently don't achieve the best success rates and accuracy. The EmoDB database was used for the experiments, and the Deep Belief Network (DBN) was fine-tuned utilizing the wake-sleep method and independent component analysis with the Fisher criterion for feature extraction. The results of this DGA-based method showed an astounding 98.5% recognition accuracy with exceptionally low false acceptance and rejection rates of only 0.01.

References

- Besacier, L., Barnard, E., Karpov, A., & Schultz, T. (2014). Automatic speech recognition for under-resourced languages: A survey. *Speech communication*, 56, 85-100.
- Jelinek, F. (1976). Continuous speech recognition by statistical methods. *Proceedings of the IEEE*, 64(4), 532-556.
- 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.
- 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.
- 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.
- Reddy, D. R. (1976). Speech recognition by machine: A review. *Proceedings of the IEEE*, 64(4), 501-531.
- 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.
- Yu, D., & Li, J. (2017). Recent progresses in deep learning based acoustic models. *IEEE/CAA Journal of automatica sinica*, 4(3), 396-409.