IJBPAS, March, 2023, 12(3): 1274-1281

ISSN: 2277–4998

International Journal of Biology, Pharmacy and Allied Sciences (IJBPAS)

'A Bridge Between Laboratory and Reader'

www.ijbpas.com

ARTIFICIAL AND DEEP KNOWLEDGE SPIRITUAL SPEAKER RECOGNITION

SELVARAJU S[1*], MANOHARAN C[2], SHASHI DEVI S[3] AND RAMACHANDRAN G[4]

1: Associate Professor, Department of Electronics and Communication Engineering, Vinayaka Mission's KirupanandaVariyar Engineering College, Vinayaka Mission's Research Foundation (Deemed to be University), Salem, Tamil Nadu, India

2: Professor, AMET Business School, AMET University, Chennai, Tamilnadu, India

3: Associate professor, Departments of Humanities and Science, Vardhaman College of Engineering, Hyderabad-501218. India

4: Assistant Professor, Department of Electronics and Communication Engineering, Vinayaka Mission's KirupanandaVariyar Engineering College, Vinayaka Mission's Research Foundation (Deemed to be University), Salem, Tamil Nadu, India

***Corresponding Author: Dr. Selvaraju S: E Mail: plcampus@gmail.com**

Received 11th May 2022; Revised 15th June 2022; Accepted 18th Aug. 2022; Available online 1st March 2023

https://doi.org/10.31032/IJBPAS/2023/12.3.6968

ABSTRACT

A method for recognising a speaker based on speech features is known as speaker recognition. Speaker recognition technology is frequently employed in a variety of fields. Most speaker identification algorithms have been developed on regular, clean recordings, but their effectiveness degrades when hearing speech with emotions. This study offers an emotional speech signals system developed using standard machine learning algorithms on an emotional speech database obtained from the University Audio visual Database of Affective Speech and Song using temporal, frequency, and spectral properties

Five models (Logistic Regression, Support Vector Machine, Variational Forest, XGBoost, and k-Nearest Neighbor) and three deep learning models (Logistic Regression, Logistic Regression, Random Forest, XGBoost, and k-Nearest Neighbor) were trained and compared in terms of performance (Long Short-Term Memory network, Multilayer Perceptron, and Convolutional Neural Network).Deep neural networks outperformed state-of-the-art models

in feelings speaker detection from voice signals after the models were evaluated. They achieved the greatest accuracy of 92 percent, exceeding machine learning techniques.

Keywords: Neural networks, machine learning, emotion recognition, speaker recognition

1. INTRODUCTION

Deep representation learning technique that mimics the operations of the human brain in processing aggregation and analysis data for intelligent applications like computer vision, speech recognition, and object detection, among others. Deep learning techniques may learn and make judgments without the need for human supervision. As a result, these techniques have sparked a lot of interest in image identification, computer vision, and speech recognition research. Deep learning technologies are used by well-known commercial businesses such as Apple, IBM, Linkedin, Oracle, Amazon, Microsoft, and Google to assist scale their business systems. Deep learning techniques have recently outperformed traditional ml algorithms like support vector machines in Data Science challenges (SVM).

Deep learning algorithms have outperformed previous approaches in terms of attitude and speaker recognition **[1-5]**. A method for recognising a user from speech is known as speaker recognition. Many factors influence speech recognition performance, including background noise, channel change, speaker, and recording quality. These factors may have a negative impact on speaker recognition by causing greater intra-speaker vocal variability, or variation between speakers. Due to the small quantity of training data available for the target language, most speaker identification algorithms are of poor quality. Low-resource environments are referred to as such. Mokgonyane *et al*. suggested a machine learning model-based speaker recognition system.

The algorithms are evaluated on clean low-resource language voice database. Using neural networks, the authors were able to achieve a 96 percent accuracy. Though there was no emotional speech in the data. Another internal element that might cause intra-speaker vocal fluctuation is emotion **[6]**. Emotion identification is the process for detecting and identifying emotions. In many cases, a speech enhancement system is constructed on a speech sound data set, and when evaluated on speech emotions data, the system performs poorly. This study offers an emotional presenter recognition system that is trained using a speech emotion data set that includes eight attitudes (disgust, surprise, fear, anger, sadness, happiness, calm, and neutral).

The data set is used to train and analyse eight different educational models using three sets of acoustics aspects of the language (Time, Frequency, and Spectral).

The following contributions are listed in alphabetical order: A survey of the literature on speech and emotion recognition is offered. A list of parameters used to educate the models is provided. We use an emotional database to train speaker recognition and achieve good results. For speaker recognition, we just provide best SVM kernel. We make a comparison of deep learning and machine learning algorithms.

We provide models for predicting speaker emotions, as well as how models differentiate between male and female speakers. The following is a summary of the paper. To begin, Section II discusses the comprehensive literature study on communicator and emotion recognition, while Section III outlines data processing, modelling (extraction and normalisation), and the methodologies used to develop and evaluate the models.

Section IV outlines the paper's debates and findings, while Section V discusses the paper's conclusion.

2. Literature Survey

The research study based on current speech enhancement systems and recognition systems is discussed in this part. A. Recognizing the Speaker: Speaker recognition performance can be influenced by a variety of factors, including the quality of the voice, the speaker's age, gender, background noise in the recording, the speaker's accent, and others. Mbogho and Katz **[7]** used two different types of Hidden Markov models to study how accent affects voice recognition quality (HMM).The first model was trained on native English speakers, while the second model was trained on impacted English speakers. When compared to models trained on native English speakers, the results showed that systems trained on affected English speakers performed better.

The quality of a speaker identification system can be measured in a variety of ways, and the performance ratings of the classification model can be difficult to comprehend, as per Ferrer *et al*. As a result, Ferrer *et al*. present a trial-based equalisation technique to apply to performance ratings in order to translate them into genuine likelihood proportions that can be interpreted perfectly probabilistically. Wu *et al*. show that data reliance can alter the sampling error of the function f when evaluating speech recognition systems. In order to train speaker identification models, neural networks were used extensively. Dropout, a well-known deep learning strategy, has shown to boost the productivity of sophisticated neural networks significantly.

The multiple connectionist models to propose a collaborative joint training technique for speech and speaker recognition systems, prompted by the

application of neural networks in speech recognition. To train a speaker recognition model, numerous features can be gathered from speech. An i-vector, for example, is a feature representation that models both the speaker and channel variability in voice signals. By expediting the extraction process at run-time, Xu *et al*. present a method for extracting the i-vector without calculating the complete posterior covariance. This is accomplished by generalising i-vector estimates, while Cumani and Laface propose e-vector, a speaker image manipulation that, like i-vectors, creates a compact representation of a speech segment.

Cumani and Laface propose e-vector, a speaker modelling technique comparable to i-vectors that creates a compact representation of a speech segment. Modipa *et al*. studied alternative strategies for acoustic modelling of a low-resource language, Sepedi, for voice recognition, whereas Manamela *et al*. **[14]** used machine learning algorithms to construct an emotion recognition system for the same language. B. Recognizing Emotions: The identification of emotions in a given speech signal is known as emotion recognition. Emotion recognition systems can be trained in a variety of ways. Deep neural networks are one of the most well-known models for developing emotion recognition systems.Vielzeuf *et al*. **[2]** offer a light deep neural network model for emotion recognition in audiovisual, despite the fact that such models are expensive to train.

The authors claimed to have achieved a state-of-the-art accuracy of 61%. Due to its strategy of attending to more relevant features that forecast the target, attention has increased the performance of deep neural networks. On speech emotion recognition, a multi-task attention based deep neural network outperforms random forest, deep neural network, and SVM approaches, according to Ma **[3]**. Using the emobase feature set **[16]**, Egorow *et al*. used a random-forest technique to identify the most essential characteristics and achieved an increase in performance using only 40 to 60% of the features Marczewski *et al*.

Propose a hybrid speech emotion recognition architecture in which a convolution neural network (CNN) is used as a feature extraction step in the first layer and a long short-term memory network (LSTM) is used as a classification layer for emotion classification using domain-specific features in the final layer. Sun *et al*. utilise CNN-LSTM to extract film character traits and SVM to classify them, whereas Wang and Hu employ SVM on enhanced Mel Frequency Cepstral Coefficients (MFCCs) to achieve state-of-the-art results. Albanie *et al*. suggest utilising a pretrained emotion detection

neural network trained on images to categorise unlabeled spoken emotion datasets.

Because noise has a detrimental effects on speaker and emotion identification systems, Pohjalainen *et al*. suggest signal denoising in the log-spectral and cepstral domains, with the authors claiming that the suggested method outperformed standard noise reduction methods.

3. Methodology

This section initially describes the gathered data, then moves on to pattern extraction and normalisation approaches, then the models, and finally, how the models are evaluated. The elements are standardized using z-score feature normalisation once they are extracted from the speech database. Finally, we utilise training data to develop models, and testing data to evaluate models by prediction and comparing them to true labels.

A. Information

A certified sensitive speech and song multimodal database, was used in this investigation, which was compiled from 24 skilled speakers (Twelve females and males) who all recorded the identical speech in a neutral India accent. Contempt, neutral, surprise, calm, afraid, angry, glad, and sad expressions are the eight emotions represented by utterances. The silence there in speech processing was erased. The voice signals were then resampled at 32000 samples per second. The songs are not included in the article. For the verbal cues of neutral and angry the spectrograms reveal that the majority of the spectral energy in a voiced section of neutral speech is concentrated at lower frequencies (in most cases below 512 Hz). The brightness activity is dispersed across a wide range of frequencies, and the rage outburst has no harmonic shape. The rage speech has an amplitude of roughly 1, whereas moderate speech possesses lower amplitude of almost 0.04.As a result, people in the English language convey emotions in speech in a variety of ways.

B. Normalization and Feature Extraction

The retrieved features and the method of normalisation utilised are discussed in this section.

1) Extraction of Features:

The acoustic characteristics of the verbal utterance characterise the speaker's identity. We utilise pyAudio Analysis **[21]** to extract the short-term features, resulting in a feature vector of size 68 that includes both the standard deviation and the mean. The Hamming window was set to a rate of 25ms and a frame size of 50ms during extraction. The 34 short-term retrieved features are classified into three domains (Frequency, Time, and Cepstral),. These characteristics are also present in **[22-24]**:

Chroma Vector, Chroma Variation, Spectral Rolloff, Chromatic Entropy,

Harmonic Flux, Resonance Centroid, and Spectral Spread are all frequency-domain properties that are founded on the size of the Discrete Wavelet Transform are cepstral-domain characteristics that are determined using an inverse DFT on the logarithmic spectrum. In emotion and speech enhancement applications, MFCCs are often exploited as acoustic aspects of speech. The following is how MFCCs are calculated:

2) Feature Normalization: This is a crucial stage in building a reliable predictive model for speech recognition system. For speaker and expression recognition systems, normalisation has been utilised.

The goal is to eliminate recordings and speech variability while maintaining speaker discrimination accuracy. We employ Sefara's z-score normalisation, which is specified by the following equation:

3) k-Nearest Relatives (kRN) is a machine-learning methodology that provides a data point's k closest neighbours to classify it. The properties of kRN are as follows: I Because it does not presume the random variable of the given data point, kRN is nonparametric. (ii) The lazy learning method is used by kRN because it generalises during the testing phase rather than the training phase. 3) During training, Random Forest creates decision trees on data samples and outputs the group that is the middle of the categories via mojority voting.

4) Aggressive Boosting is a quick and efficient solution of slope boosted decision trees that has been optimised.

5) Support Vector Machines (SVMs) are computational models with several kernels that are used to solve classification and regression problems. SVM is a discriminative model that constructs a separating hyperplane to classify a new data point. The SVM kernels listed below have been implemented.

6) A feed forward neural network is a multilayer perceptron (MLP).MLP is made up of numerous layers, each of which is activated by a separate activation function. The MLP architecture is implemented using Tensorflow. Overfitting is avoided thanks to the dropout layers. The model has 31704 parameters and is trained for different initial conditions with a batch size of 128.

7) CNNs are a sort of deep network that is commonly used in computer vision. To prevent the model from overfitting, we use Tensorflow and add dropout registration with a probability of 0.5.The model has 112024 parameters and is learned for 1000 iterations with a batch size of 128.

8) Long-term dependency networks (LSTMs) are a type of recurrent neural network that can learn long-term dependencies.

To prevent the model from overfitting, we use Tensorflow and add ejection regularisation with a probability of 0.5.The model has 156184 parameters and is developed for 1000 iterations with a batch size of 128.

5. Evaluation

The size and amount of noise in the learning algorithm, the quality of the audio recording, the type of recording device, and the type of learning approach can all have an impact on the model's quality. The measurement of how the models generalization on unseen data should be incorporated in any machine learning pipeline. There were 1296 speech samples in all. For training, testing, and assessment, we divided the data into three categories: 80%, 10%, and 10%.We used the following measurements to assess the models' prediction quality based on the test and evaluation data:

6. Results And Conversations

After training the models with accuracy, F1 score, and categorical cross-entropy, this part will cover the performance and overfitting. A. Execution After evaluating the models. We trained SVM on 4 kernels, exponential, linear, RBF, and polynomial, to illustrate the optimum SVM kernel for speaker verification using the provided features. The RBF kernel outperformed polynomial, linear, and sigmoid kernels. The performance of the Sigmoid SVM was dismal. We can see that the Sigmoid kernel failed to achieve state-of-the-art accuracy, scoring 58 percent last, followed by polynomial, linear, and RBF kernels, which scored 81 percent, 85 percent, and 88 percent, respectively. Based on these findings, the RBF kernel is appropriate for a speech recognition system, while the sigmoid is not.

Using the dataset, researchers aggregate machine simulating (LR, RF, kNN, XGBoost, SVM) and artificial neural networks (MLP, CNN, LSTM) to demonstrate the optimal technique for speaker recognition. We saw superior outcomes with RBF SVM, which has the highest accuracy, when it comes to machine learning algorithms. MLPs outperform LSTM and CNN in terms of deep learning algorithms, with a score of 92 percent. As a result, deep learning models, outperformed machine learning models in terms of F1 score and accuracy, and are thus the best models to utilise in emotional speech processing. LR also outperforms RF, kNN, XGBoost, sigmoid SVM, linear SVM, and polynomial SVM in terms of obtaining state-of-the-art outcomes.

7. Conclusion

This research described a speaker recognition system that used an emotional database of eight emotions, rather than a standard database. A data analysis on language and emotion recognition was

presented. The features were discussed, as well as feature extraction. A type of feature normalisation was described. The algorithms for learning were explained. Among different SVM kernels, the RBF kernel was shown to be suitable for speech recognition. On an emotional database of 24 speakers, deep learning algorithms outperformed machine learning techniques. Finally, we propose that this research be expanded to include I an inquiry into the identification of all the most relevant auditory features.(ii) The number of participants and speech samples is growing.

REFERENCES

[1] Z. Tang, L. Li, D. Wang, R. Vipperla, Z. Tang, L. Li, D. Wang, and R. Vipperla, "Collaborative joint training with multitask recurrent model for speech and speaker recognition," IEEE/ACM Trans. Audio, Speech and Lang. Proc., vol. 25, no. 3, pp. 493–504, Mar. 2017.

[2] V. Vielzeuf, C. Kervadec, S. Pateux, A. Lechervy, and F. Jurie, "An occam's razor view on learning audiovisual emotion recognition with small training sets," in Proceedings of the 20th ACM International Conference on Multimodal Interaction, ser. ICMI '18. Boulder, CO, USA: ACM, 2018, pp. 589–593. [Online]. Available: http://doi.acm.org/10.1145/3242969.3264980

[3] F. Ma, W. Gu, W. Zhang, S. Ni, S.-L. Huang, and L. Zhang, "Speech emotion recognition via attention-based DNN from multi-task learning," in Proceedings of the 16th ACM Conference on Embedded Networked Sensor Systems, ser. SenSys '18. Shenzhen, China: ACM, 2018, pp. 363–364.

[4] B. Sun, Q. Wei, L. Li, Q. Xu, J. He, and L. Yu, "LSTM for dynamic emotion and group emotion recognition in the wild," in Proceedings of the 18th ACM International Conference on Multimodal Interaction, ser. ICMI '16. Tokyo, Japan: ACM, 2016, pp. 451–457.

[5] S. Albanie, A. Nagrani, A. Vedaldi, and A. Zisserman, "Emotion recognition in speech using cross-modal transfer in the wild," in Proceedings of the 26th ACM International Conference on Multimedia, ser. MM '18. Seoul, Republic of Korea: ACM, 2018, pp. 292–301.

[6] N. K. Mudaliar, K. Hegde, A. Ramesh and V. Patil, "Visual Speech Recognition: A Deep Learning Approach," *2020 5th International Conference on Communication and Electronics Systems (ICCES)*, 2020, pp. 1218-1221, doi: 10.1109/ICCES48766.2020.9137926.