Identification of Speech Signal in Moving Objects using Artificial Neural Network System

DIWAKAR BHARDWAJ[1], RAKESH KUMAR GALAV[2]

[1]*Department of Computer Engineering and Application*
GLA UNIVERSITY, MATHURA
diwakar.bhardwaj@gla.ac.in

[2]*Department of Computer Engineering and Application*
GLA UNIVERSITY, MATHURA
rakesh.kumar@gla.ac.in

Abstract: The speech signal moving objects regarding the speaker's personality. A speaker recognition field is about retrieving the name of the individual voicing the speech. The effectiveness of accurately identifying a speaker is focused solely on vocal features, as voice contact with machines is becoming more prevalent in tasks like telephone, banking transactions, and the transformation of data from speech databases. This review illustrates the detection of text-dependent speakers, which identifies a single speaker from a known population. The program asks the user to utter voice. Program recognizes the person through evaluating the voice utterance codebook with the voice utterance codebook held in the database and records that may have provided the voice speech. Furthermore, the features are removed; the speech signal is registered for 6 speakers. Extraction of the function is achieved using LPC coefficients, AMDF calculation and DFT. By adding certain features as input data, the neural network is equipped. For further comparison the characteristics are stored in models. The characteristics that need to be defined for the speakers were obtained and analyzed using Back Propagation Algorithm to a template image. Now this framework trained correlates to the outcome; the source is the characteristics retrieved from the speaker to be described. The weight adjustment is done by the system, and the similarity score is discovered to recognize the speaker. The number of iterations needed for achieving the goal determines the efficiency of the network.

Keywords— Speech Recognition, Artificial Neural Network, Lib ROSA feature

1. INTRODUCTION

Signal processing refers to the processing of the frequency without worries for the reliability or quality of such channel estimation. In appreciation of expression, speech is usually processed on even a frame-by - frame premise with only the worries that perhaps the framing would be either speech or mute the accessible recorded signal can be identified as frame of voice containing high data content than useless frames with relating to a specific purpose. To classify available speech clips, we have been exploring a speaker recognition system. Instead, using a specific approach, we identify a framework for defining certain frames as available. Understanding how accurate the information is in a speech environment can, nevertheless, be very helpful and useful. It is where identification and extraction of functional speech will play a vital role. The accessible speech frames can be classified as speech frames which showed increased statistical validity than inaccessible images with respect to a specific function. In order to classify functional speech frames, we examined a speaker recognition scheme. We then developed a framework for defining blocks like this functional using a different methodology.

Paradigms of Speech Recognition

Speech processing-Recognize which one of the target group spoke a specific speech.
Speech verification-Ensure that the speaker in question is the one he appears to be. Program asks the user to include an ID, who appears to become the speaker. Program confirms user by matching codebook with user- specified speech utterance. Unless it coincides with the defined threshold therefore the recipient's identity argument is otherwise denied.

1. Identity of the speaker-identifies a given speaker of either a known population. The program asks the user to express voice. System recognizes the person by combining the voice utterance codebook with the voice utterance codebook held in the database and lists that may have provided the voice word. There are many two forms of defining speakers

- Text dependent – spoken word corresponds to known text, cooperative user, applications type 'PIN'
- Free text – no limitations on what the speaker are saying, possibly unresponsive users

2. Reaches for Identification Statement

The Acoustic Phonetic Approach

The acoustic phonetic method is based on the idea of auditory phonetics whether postulates that a collection of fixed, recognizable phonetic classes exists in the speech Language but that the phonetic units are generally defined by a set of attributes which can be seen across period in the voice signal and its range. While the functional aspects the phonology functions are extremely variable with both the speaker and the adjacent phonetic systems, the laws governing that variability are believed to be fast paced and can be readily learned and applied in realistic circumstances. Then the first phase in the approach is called the process of segmentation and classifying. This involves segmenting the speech signal into distinct (in Time) areas where signal's acoustic properties are indicative of one of many phonetic units or groups, but instead adding any or even more phonetic marks as per acoustic properties from each segmented image. A secondary process is intended for voice recognition. This step two tries to assess a specific word (or a string of words) from of the phonetic label sequence generated during the first phase, which would be compatible from the limitations of the task of speech recognition.

The Pattern Recognition Approach

The Pattern Recognition method to voice is essentially one where speech signals are used explicitly without specific identification of features (during this acoustic-phonic context) and differentiation. As for most pattern recognition methods, the system has two stages – namely, speech pattern preparation, and pattern recognition by pattern comparing. Speech is implemented into a process through a training process The idea is that if appropriate variations of a template are included in test data given for the algorithm (whether it sounds a word, a phrase, etc.), the training process will be willing to do so properly describe the acoustic impedance of the model without regard to or awareness of any pattern provided to the train. This type of speech characteristics through training is named the classifiers of patterns. Here the computer learns that voice class functional aspects are accurate yet replicable through plenty of learning objects of the sequence. The usefulness of this approach is the pattern matching phase with any possible pattern acquired during the training process and the classification of the unidentified speech as per the reliability of the template matching

Benefits of Detection in Design technique

Easy to use. The approach is fairly easy to comprehend. It is rich in justification for the numerical as well as the Social Concept discovery questions used during training and decrypting. It is better defined and commonly utilized.

Robustness and invariance to specific language, user and functionality sets algorithms and decision rules for the pattern comparison. This property such as high strength the equation ideal for a wide variety of speech units, word vocabulary, populations of speakers, context settings, conditions of transmission etc.

Good Quality Proven. The design identification method to speech recognition reliably offers high output on any challenge that is technically appropriate and provides a straightforward roadmap for expanding the technology in a wide variety of ways.

A. The Artificial Intelligence Approach

The artificial intelligence methodology to voice is a mixed approach to acoustic phonetics and pattern recognition, wherein concepts and principles of both approaches are used. The artificial intelligence approach aims to computerize the process of recognition as per the manner in which an individual applies information in analysis, evaluating and eventually making a judgment on the acoustic characteristics measured. Use of an expert framework for segmentation and marking is especially among the strategies used in the class of methods. Availability of neural networking may represent a separate computational approach to voice recognition, or could be viewed as an application of integration that could be integrated into any of the classical approaches listed above.

3. LITERATUREREVIEW

Hossein Salehghaffari designed a prototype coevolutionary neural network topology for speech recognition to collect and dump presenter and non-presenter details simultaneously, simultaneously. In the training process, the network is taught to differentiate between different personalities of speakers to establish the context model. Some of the key aspects is making the templates for speakers. All of the preceding methods construct speaker projections based on the context model averaging the speaker depictions. We solve this issue using the Siamese paradigm to further fine-tune the trained model to create a racially discriminatory feature vectors to differentiate between the same speakers and different speakers irrespective of their gender. They provide a framework that collects the speaker-related information simultaneously and generates reliability to variations within the speaker. This shown that the method proposed outperforms the conventional methods of validation which generate speaker models from the context model [1].

In the area of speaker recognition, Dávid Sztahós summarized the deep learning techniques applied, both in testing and detection. Speaker identification was a commonly used subject in the field of speech technology. A lot of theoretical work has been undertaken and no progress has been made over the last 5-6 years. Nevertheless, as deep learning approaches develop in most areas of machine learning, they are also being replaced with the former state-of-the-art methods of speech recognition. It would seem that DL is becoming the answer to state- of-the-art for authentication and recognition of speakers. For most novel works the regular x-vectors, for addition to I are used as the basis. The growing quantity of data collected opens the territories to DL, where they're more effective [2].

Yanick Lukic 's research on Deep learning, especially in the context of Coevolutionary neural networks (CNNs), has in recent years triggered significant improvements in computer vision and related fields. This advance is due to the change from developing applications and ascending scale sub-systems to learning features and end-to - end recognition systems from almost raw information. Nevertheless, It's out there popular to use handmade processing chains, such as MFCC features and GMM-based models, for speakers cluster analysis. In this paper they using basic spectrograms as inputs to a CNN and research the optimum solution for speaker recognition and clustering of those networks. We also expand on the issue of how to move a platform, qualified to recognize speakers, to cluster speakers. A very well-known one TIMIT dataset, we illustrate our method, obtaining comparable results the state of the art – except no handmade applications. [3] [English version]

Chularat Tanprasert implemented a text-dependent speaker recognition method for the Thai language based on the neural networks. Linear Prediction Coefficients (LPC) are obtained from voice signal and the vectors generated by the function. These images are useful into a neural network with multilayer perceptron's (MLP) with a back-propagation training algorithm for learning and process recognition.

A.Rajeswari, In this paper,A model of emotional speech recognition algorithm using Tamil language is suggested using Linear Predictive Coding (LPC) and Parameters dependent approach to diagnosis attempts at depression and suicide. The framework is equipped with multilayer perceptron back propagation approach utilizing Artificial Neural Network (ANN). The emotional speech inventory is retrieved and LPC-based tests are recorded, and parameters such as strength, average value-based recognition method. LPC identification is simple and reliable and 90 percent is the highest acknowledgement rate reached for LPC dependent acknowledgement.

Anurag Bajpai1, the paper explains how to effectively build an Automated Speech Recognition (ASR) framework for universal control. The design is based on an algorithm which extracts isolated words from a continuous signal of expression. The voiced component of speech signal extraction function is performed by Mel Frequency Cepstral Coefficients (MFCC) while Artificial Neural Network (ANN) is used for processing and patterning. The enhanced dismissal of unwanted speech commands is obtained by the Euclidean distance- based judgments, calculated between qualified and validated voiced commands. Also enhanced is the signal SNR, at the pre-processing level.

4. PROPOSED SYSTEMARCHITECTURE

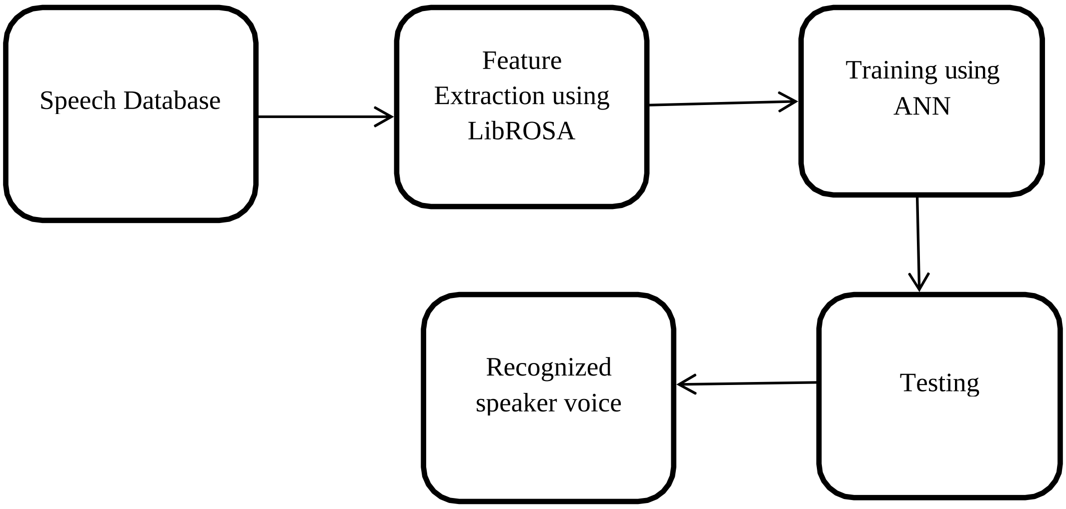


Fig.1 System Architecture of Proposed model

5. METHODOLOGY

Artificial Neural Networks can be represented best as weighted directed graphs, in which the nodes are created by artificial neurons and the relationship between neuronal outputs and neuronal inputs can be expressed by weigh directed edges. The Artificial Neural Network generates the feedback signal from the outer environment as a result of a pattern and image formed by vectors. These inputs are then mathematically described by the notations x(n) in order to increase n number of inputs. Every input is then multiplied by its associated weights. Such weights usually represent the degree of interconnection between neurons inside the artificial neural network, in simple terms. Both weighted inputs are summed up within the computing unit.

If the correct choice is incorrect, a bias may be added to make the output non-zero, or to scale it up to the device's outcome. Bias has the weight and requests a comparison to 1. There the number of weighted inputs will be in the range 0 to positive infinity. To maintain the outcome inside the goal value limits a certain reference amount is benchmarked. And instead, the number of weighted inputs is transferred through the activation function. Typically, to transfer functions that are needed to get the necessary output, the activating functionality is set. The activation function has various forms of features, but often only linear or non-linear sets. The Activation functions Discrete, Sigmoidal (linear) and Tan hyperbolic sigmoidal (nonlinear) are among the most commonly known classes of activation functions.

Input layer:

The layers of data comprise the artificial neurons for collecting input from outside. It is where the real testing takes place on the network, or comprehension occurs so it functions.

Output layer:

The output layers contain units that respond to the data entered into the system, as well as whether or not any task has been completed.

Hidden layer:

The hidden layers in between input layers and output layers are described unknown. The only task for a hidden layer is to transform the information it in to something usable that the output layer / unit will use in some way. Many artificial neural networks are always interconnected, which ensures that every one of the hidden layers is directly linked to the neurons in their input layer and even to the output layer. It allows for a full learning process and learning also happens to the maximum and with each iteration the weights within the artificial neural network are increasing.

6. EXPERIMENTALSTUDY

The following are the steps to implement the proposed methodology of speaker recognition:

Step 1: Collection of voice datasets from different persons or actors for differential voice collection. The figure below shows the voices folders:

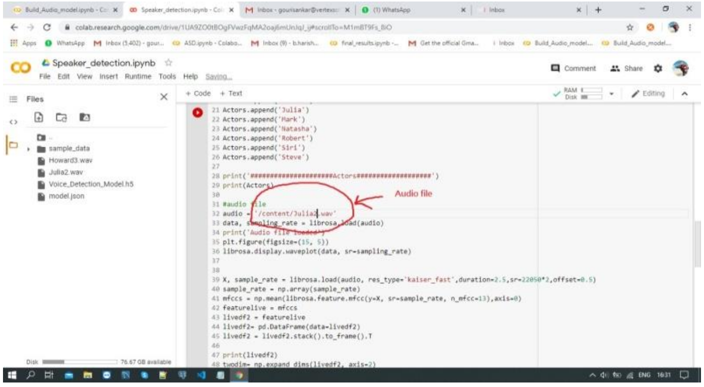
Step 2: Feature Extraction using Librosa python tool

Step 3: Divide the dataset into test and train parts that is in 80:20 ratio. Step 4: Set labels or the voices owners names for Recognition

Step 5: Training the ANN model for about 700 epochs for better prediction.

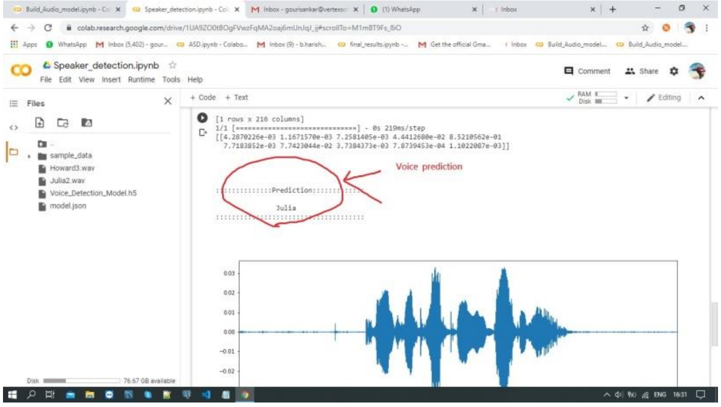
Step 6: Save the model as Voice_detection.h5 and the configuration json file i.emodel.json Now we can use the saved model to recognize voices.

7. RESULTS



```
Actors.append('Julia')
Actors.append('Mark')
Actors.append('Natasha')
Actors.append('Robert')
Actors.append('Siri')
Actors.append('Steve')

print('####################Actors####################')
print(Actors)

#audio file
audio = '/content/Julia2.wav'
data, sampling_rate = librosa.load(audio)
print('Audio file loaded')
plt.figure(figsize=(15, 5))
librosa.display.waveplot(data, sr=sampling_rate)


X, sample_rate = librosa.load(audio, res_type='kaiser_fast',duration=2.5,sr=22050*2,offset=0.5)
sample_rate = np.array(sample_rate)
mfccs = np.mean(librosa.feature.mfcc(y=X, sr=sample_rate, n_mfcc=13),axis=0)
featurelive = mfccs
livedf2 = featurelive
livedf2= pd.DataFrame(data=livedf2)
livedf2 = livedf2.stack().to_frame().T

print(livedf2)
twodim= np.expand_dims(livedf2, axis=2)
```

Fig.2 Set the audio file for speaker prediction



```
[1 rows x 216 columns]
1/1 [==============================] - 0s 219ms/step
[[4.2870226e-03 1.1671570e-03 7.2581405e-03 4.4412680e-02 8.5210562e-01
  7.7183852e-03 7.7423044e-02 3.7384373e-03 7.8739453e-04 1.1022087e-03]]


::::::::::::::Prediction::::::::::::::

                Julia
::::::::::::::::::::::::::::::::::::::
```

Fig.3 speaker predicted as Julia

8. CONCLUSION

The obtained results indicated that the variation in the tests is small for the same speaker spoken at different instances. The program works perfectly to distinguish speaker from different inputs. Small vocabulary, and limited number of users. The algorithm only functions with Artificial Neural Network files called '.wav.'s must be well trained. On our trained model we got an accuracy of more than 97% as we trained with around 600 voices from 10 different persons where 5 from male and 5 from female voices.

REFERENCES

1. HosseinSalehghaffari "*Speaker Verification using Convolutional Neural Networks*" NYU Tandon School of Engineering (Polytechnic Institute), NY 11201, USA
2. DávidSztahó, *"Deep learning methods in speaker recognition: a review*" Department of Telecommunication and Media Informatics, Budapest University of Technology and Economics, Budapest, Hungary
3. YanickLukic, Carlo Vogt, Oliver Durr "*speaker identification and clustering using convolutional neural networks*", 2016 ieee international workshop on machine learning for signal processing
4. ChularatTanprasert, "*Text-dependent Speaker Identification Using Neural Network On Distinctive Thai Tone Marks*",Software and Language Engineering Laboratory
5. AmnaIrum and Ahmad Salman, "*Speaker Verification Using Deep Neural Networks: A Review*" International Journal of Machine Learning and Computing, Vol. 9, No. 1, February 2019

6. A. Reynolds, T. F. Quatieri, and R. B. Dunn, "*Speaker verification using adapted Gaussian mixture models*," Digital Signal Processing, vol. 10, pp. 19-41,2000.
7. D. A. Reynolds and R. C. Rose, "*Robust text-independent speaker identification using Gaussian mixture speaker models*," IEEE Trans. on Speech and Audio Processing, vol. 3, pp. 72-83,1995
8. S. Tranter and D. Reynolds, "*An overview of automatic speaker diarisation systems*," IEEE Trans. Audio Speech Lang. Process., vol. 14, no. 5, pp. 1557-1565,2006.
9. A. Salman and K. Chen, "*Exploring speaker-specific characteristics with deep learning*," in Proc. IJCNN, pp. 103-110,2011.
10. T. Yamada, L. Wang, and A. Kai, "*Improvement of distant-talking speaker identification using bottleneck features of dnn*," in Proc. INTERSPEECH, pp. 3661–3664,2013.
11. J. Kulshrestha and M. K. Mishra, "An Adaptive Energy Balanced and Energy Efficient Approach for Data Gathering in Wireless Sensor Networks", Ad Hoc Networks Journal, Elsevier, Vol. 54, pp. 130-146, 1 November 2016 [SCI. Impact Factor: 3.643].
12. M. Kumar and C. Bhatnagar, "Hybrid tracking model and GSLM based neural network for crowd behavior recognition", Journal of Central South University, Springer, Vol. 24, No. 9, pp 2071 - 2081, 7 October 2017 [SCI, Impact Factor: 1.249].
13. M. Kumar and C. Bhatnagar, "Crowd Behaviour Recognition using Hybrid Tracking Model and Genetic Algorithm Enabled Neural Network", International Journal of Computational Intelligence Systems, Atlantis Press,Vol. 10, Issue 1, pp 234 - 246, 1 January 2017 [SCI, Impact Factor: 1.14]
14. J. Kulshrestha and M. K. Mishra. "Energy balanced data gathering approaches in wireless sensor networks using mixed-hop communication." Computing (2018), Springer, Vol. 100, pp. 1033-1058, 20 March 2018 [SCI Impact Factor: 1.589]
15. Varun K L Srivastava, N. Chandra Sekhar Reddy, Dr. Anubha Shrivastava, "An Effective Code Metrics for Evaluation of Protected Parameters in Database Applications", International Journal of Advanced Trends in Computer Science and Engineering, Volume 8, No.1.3, 2019. doi.org/10.30534/ijatcse/2019/1681.32019
16. Prasad, K.S., Reddy, N.C.S. & Puneeth, B.N. A Framework for Diagnosing Kidney Disease in Diabetes Patients Using Classification Algorithms. SN COMPUT. SCI. 1, 101 (2020).
17. Kumar, R., Bhardwaj, D., Mishra, M.K. 2020. Enhance the Lifespan of Underwater Sensor Network through Energy Efficient Hybrid Data Communication Scheme 2020 International Conference on Power Electronics and IoT Applications in Renewable Energy and its Control, PARC 2020 9087026, pp. 355-359
18. Kumar, R., Bhardwaj, D. 2020. An improved moth-flame optimization algorithm-based clustering algorithm for VANETs Test Engineering and Management 82(1-2), pp. 27-35
19. Bhardwaj, D., Chaturvedi, A. 2020. A Hybrid Resource Optimization Technique using Improved Fuzzy Logic Guided Genetic Algorithm for 5G VANETs Test Engineering and Management 82(1-2), pp. 36-44
20. Kumar, M., Bhardwaj, D.2019 Optimized cluster head and secret key comparison based secure routing in WSN Journal of Advanced Research in Dynamical and Control Systems 11(11 Special Issue), pp. 183-188.