IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Issue-11 E-ISSN: 2347-2693

Neural Network Based Speaker Verification using GFCC

Sukhandeep Kaur[1]* and Kanwalvir Singh Dhindsa[2]

Dept.of CSE, BBSBEC Fatehgarh Sahib, Punjab Technical University,India
Sukhansandhu786@gmail.com, kanwalvir.singh@bbsbec.ac.in

www.ijcseonline.org

Received: Oct/22/2015 Revised: Nov /04/2015 Accepted: Nov/17/2015 Published: Nov/30/2015

Abstract - Speaker confirmation is feasible method of controlling access to computer and communication networks. Speakers resonance is different due to physiological differences such as vocal tract size, larynx size and other voice produce organs, and speaking manner differences such as accent and often used words. The task of automatic speaker identification is to identify the underlying speaker or confirm the claimed speaker from a sound recording, by exploiting these differences. This paper introduce the important concepts of speaker confirmation for security system.

Keywords- Gaussian Mixture Model, Finite Impulse Response, Artificial Neural Network, Gaussian

1. INTRODUCTION

Voice is the primary way of communication between humans. Speaker identification is the process of automatically recognizing an individual on the basis of characteristics of words spoken. Speaker identification has all the time target on security system for the management, the access to protected information from used by someone. Speaker confirmation is the branch of biometric certification. In addition, speaker recognition is considered into text-dependent and text-independent recognition based on whether to imagine the knowledge of written text. Speaker features encode speaker particular characteristics, and are extracted from time domain signals. Generally used speaker features contain short-time spectral/cepstral features, spectro-temporal features, prosodic features, etc. Short-time features are generally derived from short-time Fourier transform (STFT). The current usually used methods for speaker identification are GMM (Gaussian Mixture Model) , HMM (Hidden Markov Model), ANN (Artificial Neural Network) etc. GMM continue of Gaussian probability density function working well in speaker recognition systems because of its capacity to approximate the probability density distribution of arbitrary shape perfectly. HMM performs well in speaker recognition has a high accuracy. ANN is a computational model based on the structure and functions of biological neural networks. ANN has three layers which are interconnected. The first layer consists of input neurons. Those neurons send data on to the second layer, which in turn sends the output neurons to the third layer.

2. LITERATURE REVIEW

Mukherjee et al. [1] discussed voice is one of the most assure and develop biometric modalities for access control. This paper presents a new method to recognize speakers by involve a new set of characters and using Gaussian mixture models (GMMs). In this research, the method of shifted MFCC was introduced so as to incorporate accent information in the recognition algorithm. Wang and Ching [7] focussed on the features estimation method leads to robust recognition performance, specially at low signal-to-noise ratios. In the context of Gaussian mixture model-based speaker recognition with the presence of additive white Gaussian noise, the new approach produces logical reduction of both recognition error rate and equal error rate at signal-to-noise ratios ranging from 0 to 15 db. Faraj and Bigun [8] presented the first extended study investigation the added value of lip motion features for speaker and speech-recognition applications. Digit identification and person-recognition and confirmation experiments were conducted on the publicly available XM2VTS database showing good results.

Sinith et al. [9] detailed the lay accent on text-Independent speaker recognition system where we adopted Mel-Frequency Cepstral Coefficients (MFCC) as the speaker speech feature argument in the system and the concept of Gaussian Mixture Modeling (GMM) for modeling the extracted speech feature. The Maximum likelihood ratio detector algorithm is used for the decision making process. Agrawal et al.[5] discussed about the prosodic features based text dependent speaker recognition where the prosodic features can be extracted through linear predictive coding. Formantsare efficient parameters to characterize a speaker’s voice. Formants arecombined with their corresponding amplitudes, fundamental frequency, duration of speech utterance and energy of the windowed section. This feature vector is input to machine learning (ML) algorithms for recognition.

Dutta [7] explained a new approach to text dependent speaker identification using the complex patterns of variation in frequency and amplitude with time while an individual utters a given word through spectrogram segmentation and template matching. The optimally segmented spectrograms are used as a database to successfully identify the unknown individual from his/her voice.

3. PROBLEM FORMULATION

Verification procedure has been enforced on many companies in speaker verification field . The problem arises in this type of communication if there is a noise in the medium or some environmental noise on speakers end, which degrades the performance of these systems. Therefore this idea is used as a motivation to work on robust speaker recognition system. After surveying the literature, it is found that gammatone filter-bank cepstral coefficients gives good performance in noisy conditions than their other alternatives i.e. mel-frequency cepstral coefficients, linear predictive coding, etc. In this work GMM's is used to reduce the feature space in order to use them in ANN training and testing purposes.

4. STAGES IN SPEAKER RECOGNITION SYSTEM

The various stages for speaker recognition are :

1. Input - Text dependent speech is taken as a input signal. Original signals are recorded.
   (i) Pre-emphasize – Done by FIR filter. It removes local noise and make smooth frequency
   (ii) Feature extraction – Done by using GFCC, single delta and double delta of GFCC.
2. Feature Extraction is done by GMM.
3. Artificial Neural Network is used for testing and training.

5. BACK PROPAGATION ALGORITHM

Many Neural Network types have been proposed over the years. Back Propagation is the training or learning algorithm rather than the network itself. The process of determining the error rates of each neuron that impact the output is called back propagation. The neurons of the input layer are fully connected to the hidden layer and the outputs of the hidden layer are fully connected to the output layer.

The instruction using Back Propagation Algorithm involves four stages [38]:

1. **Initialization of weights**- It is general training to assign, at random generated positive and negative quantities as the initial weight values.
2. **Feed forward**- First part is getting the values of the hidden layer nodes and second part is using those values from hidden layer to calculate the values of output layer.
3. **Back Propagation of errors**- It passes error signal backward throughout the network during training to update the weights of networks.
4. **Updating of weights**- After error calculation weights are updated.

6. RESULTS

Voice is recorded as input speech signal. Pre-emphasis is implemented by using FIR filter. Feature extraction is in terms of GFCC's and production of deltas and double-delta features. Feature space and Normalization is decreasing by using GMM modeling. Back propagation algorithm has been used for learning of the neural network After instruction the neural network is tested for the all test feature vector.

Steps for performing speaker identification are as shown in Fig.1.

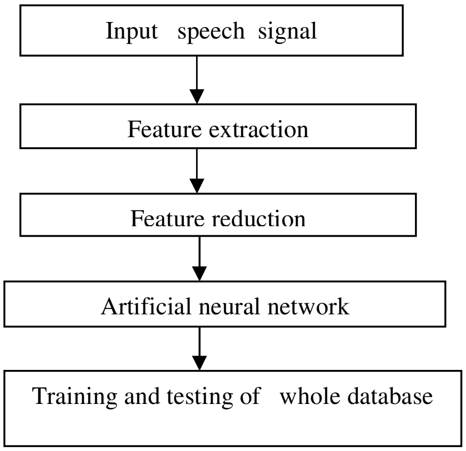
Input speech signal
Feature extraction
Feature reduction
Artificial neural network
Training and testing of whole database

Fig.1: Steps of speaker recognition system

5.1 Sensitivity and specificity

The sensitivity of a classifier is the fraction of the voice samples correctly classified as that specific student class

$$Se = \frac{TP}{TP + FN}$$ ......Eq.(1)

The specificity is the fraction of normal speech samples correctly classified as normal class.

$$Sp = \frac{TP}{TP + FP}$$ ....Eq.(2)

Sensitivity and specificity of all speakers is illustrated below in Fig.2.

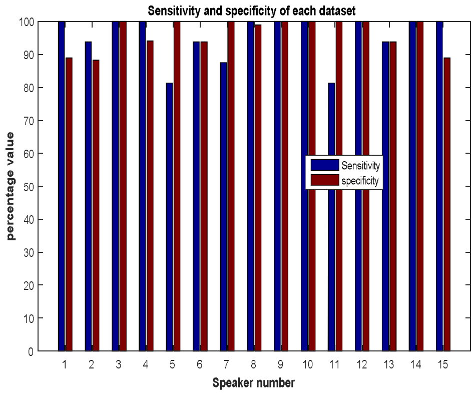

Fig.2: Sensitivity and specificity

In the above Fig.2:

At speaker 1- There are total 16 samples which are checked with trained 16 neural network, so its sensitivity is 100.

- But specificity is 88.9, because it shows 2 false positive results in which two more samples are matching with it.
- At speaker 3 – Both sensitivity and specificity are 100% because there is no false positive or false negative result.

5.2 Accuracy - Mean of specificity and sensitivity

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad .....Eq.(3)$$

Accuracy of all speakers is illustrated below in Fig.3.

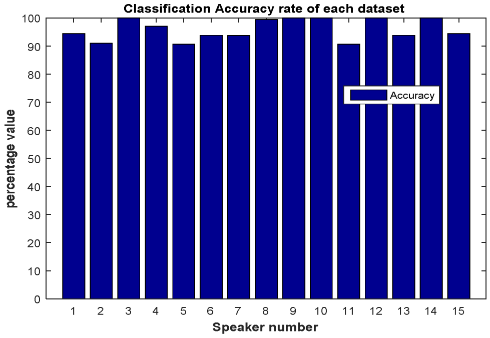

Fig.3: Accuracy values

In the above Fig.3:

- At speaker 1- sensitivity is 100 but its specificity is 88.9, after taking its mean it shows its accuracy is 94.5 .
- At speaker 3- Both of its sensitivity and specificity is 100, then it shows its accuracy 100

7. CONCLUSION AND FUTURE SCOPE

In order to confirm and identify the input speaker, its audio features which are extracted using 'gammatone' filter bank. To get the classified results in last stage speech recognition systems, particular characteristics of the spectral estimated are required. Hence these are given by the combination of GFCC, single deltas and double delta GFCC's. The dataset consist of 15 persons in which every person speaks four words four times each. For categorization or verification process, GMM models are created to produce final features which are then fed to artificial neural networks. From GFCC's, delta GFCCs and double delta GFCC's three models are generated using a particular Gaussian mixture set. The results detailed that the proposed model gives 95 % verification rates. It has been found that this method identifies those speakers as well who speaks with constraint as we have taken data from constrained environments in which each speaker was chosen at random without selecting a particular origin, dialect, language, gander, age etc.

In future, work text independent data can be used for recognition of individuals. Also, other feature extraction and classification techniques can be implemented and compared.

References

[1] R.Mukherje, I.Tanmoy, and R.Sankar, "Text dependent speaker recognition using shifted MFCC". Southeast on, 2013, Proceedings of IEEE, Orlando, FL, USA,Vol.9, pp.1-4.

[2] Wu Ju, "Speaker Recognition System Based on Mfcc and Schmm". Symposium on ICT and Energy Efficiency and workshop on Information Theory and Security, 2005, Dublin Ireland, pp. 88 – 92.

[3] D.A. Reynolds, "Speaker Identification and Verification Using Gaussian Mixture Speaker Models", Speech Communication, Vol.17,1995, No. 1-2, pp. 91-108.

[4] W.Junqin, and Y. Junjun, "An Improved Arithmetic of Mfcc in Speech Recognitions System". Electronics, Communications and Control (ICECC), International Conference on. IEEE, 2011, Zhejiang China, pp .719-722.

[5] U. Shrawankar, and V.M. Thakare, "Techniques for Feature Extraction In Speech Recognition System: A Comparative Study."International Journal Of Computer Applications In Engineering, Technology And Sciences, Vol. 2, No.5,2010, pp. 412-418.

[6] H.Hermansky, "Perceptual Linear Predictive (PLP) Analysis of Speech". Speech Technology Laboratory, Division of Panasonic Technologies,Vol.87,No.4, 1990,pp. 1738-1752.

[7] N. Wang, and P.C. Ching, "Robust Speaker Recognition Using Denoised Vocal Source and Vocal Tract Features speaker verification", IEEE Transaction on Audio Speech and Language processing, Vol. 19, No. 1 ,2011, pp. 196-205.

[8] M.I. Faraj, and J. Bigun,"Synergy of lip-motion and acoustic features in biometric speech and speaker recognition".Computers, IEEE Transactions on computers Vol.56, No.9,2007, pp. 1169-1175.

[9] M.S. Sinith, A.Salim, K. Gowri Shankar ,S. Narayanan, and V. Soman, "A novel method for Text-Independent speaker identification using MFCC and GMM".Audio Language and Image Processing (ICALIP), International Conference on. IEEE,Shanghai, 2010,Vol.5, pp.292-296.

[10] A.Solomon off,. "Channel compensation for SVM speaker recognition". Odyssey.Vol. 4,2004, pp.57-62.

[11] R. Collobert, and S.Bengio, "SVM Torch: Support vector machines for large-scale regression problems". The Journal of Machine Learning Research, No .1,2001, pp. 143-160.

[12] D.E.Sturim, and D.A. Reynolds, "Speaker Adaptive Cohort Selection for Tnorm in Text-Independent Speaker Verification."ICASSP,No.1,USA ,2005, pp.741-744.

[13] G.S.V.S.Sivaram, Thomas, and H.Hermansky, "Mixture of Auto-Associative Neural Networks for Speaker Verification". INTERSPEECH, Baltimore, USA,2011, pp. 2381-2384.

[14] S.Gfroerer, "Auditory instrumental forensic speaker recognition". Proceedings of Eurospeech,Geneva, 2003,pp. 705–708.

[15] H.R.Bolt, and F.S.Cooper, "Identification of a Speaker by Speech Spectrograms", American Association for the Advancement in Science, Science, Vol. 166, 1969.pp. 338–344.

[16] D.Charlet, D.Jouvet, and O.Collin, "An Alternative Normalization Scheme in HMM-based Text-dependent Speaker Verification", Speech Communication, Vol. 31,2000, pp. 113-20.

[17] T.Dutta, "Dynamic Time Warping Based Approach to Text-Dependent Speaker Identification Using Spectrograms," Congress on Image and Signal Processing, Vol. 2, No.8 ,2008, pp. 354-60.