Language and Text Independent Speaker Recognition System using Artificial Neural Networks and Fuzzy Logic

J Sirisha Devi

***Abstract:** In this era of technological advancement, the new techniques are evolving for better man-machine interaction. Initially, this urge of interacting with machines was the main reason behind invention of input-output devices like keyboard, mouse, monitor, printer, joystick, scanner, touch-screens, and trackball etc. However, none of these above said inventions are able to provide verbal interaction of human and machines, which is the natural means of communication for many centuries. This lack of communication with machines using speech leads the researchers towards inventing the speech processing systems for better human machine interaction using speech signals. In the present paper, the performance of an algorithm for language and text independent speaker recognition systems based on fuzzy logic and ANNs is evaluated. The efficiency of speaker recognition system with noisy speech samples of user defined database is higher than that of TIMIT database.*

***Index Terms:** Speaker Recognition, artificial Neural Networks, Fuzzy Logic.*

I. INTRODUCTION

Speech recognition by machine may be defined as the conversion of human speech signal into textual form automatically by the machine without any human intervention, providing a transcription\interpretation of what a human is speaking while machine listen. Speech recognition means identification of the spoken sentences\words by a machine. Then, these sentences\words are converted into a format understandable by machine, then, compare this with a previously stored template\dictionary of identified words. It is one of the tasks computer outperforms humans. Scientists have done regress research on the human ability of recognizing and discriminating speech signals. By determining factors detailing speaker specific information, researchers have designed reliable speaker recognition systems. In this era of digital computers, researchers have developed such automatic speaker recognition systems that could outperform human listeners on similar task (Rosenberg, 1973). Many limitation and challenging problems remain to be overcome with automatic speaker recognition systems. Speech signal retains information at several levels. Primarily, speech conveys message intended to deliver through words being spoken. But secondarily, speech also has information about the speaker. Speech recognition is related to the extraction of the linguistic message in the uttered speech while speaker recognition is identifying a person who is speaking. Speaker recognition is a stream of biometric authorization which deals with the automatic identification of individual person using some inherent characteristics of that individual. Biometric is a branch of science which studies elements of the life of humans, animal's and\or plants. This is a Greek word where 'Bio' means life and 'Metric' means measure. The process of identification or recognition of the identity of any person on the basis of one's physiological\behavioral characteristics is called Biometrics [1]. Biometric authorization has been very crucial method for human and machine interaction systems for specific tasks with security concerns. Besides speech, several physical and behavioral patterns, like eyes, face, fingerprint, signature, etc., are available for biometric authorization [2]. In fact, selecting a good and robust biometric pattern for biometric authorization should have consideration of characteristics like: (1) Robust, (2) Distinctive, (3) Accessible, and (4) Acceptable.

Outline of the remaining paper, section -2 gives a brief review on research work so far done for speaker recognition. Section -3 presents the proposed novel algorithm. Section-4 gives details on experimentation and results obtained. Section -5 presents the conclusion.

II. RELATED WORKS ON SPEAKER RECOGNITION

Speaker recognition is a technology which is used to authenticate persons from their speech samples [3]. A standout amongst the most vital difficulties in speaker acknowledgment originates from irregularities in the distinctive kinds of discourse tests and their quality. One such issue, which has been the prime focal point of specialists, is the issue of channel bungle, in which the discourse information has been gathered utilizing one mechanical assembly and the test has been recorded by an alternate channel [4]. Note that the wellsprings of jumble shift and are for the most part very confounded. There could be any blend and more often than not will be not constrained to jumble in the handset or recording contraption, the system limit and quality, commotion conditions, sickness related conditions, push related conditions, change between various media, and so on. A few methodologies include standardization or the like to either change the information (crude or in the element space) or to change the model parameters [5].

Revised Manuscript Received on 30 March 2019.
* Correspondence Author

J Sirisha Devi*, Department of Computer Science and Engineering
Institute of Aeronautical Engineering, JNTU (H) India.

In forensic speaker recognition various other factors must be considered because there are several additional problems as compared to biometric speaker recognition. The main cause of these problems is mostly the uncontrolled recording conditions and small amount of available background speech data for comparison. To develop a database for forensic application, developer has to abide by some conditions laid down by the law officials [6]. For example, a forensic database developed by Netherlands Forensic Institute known as Forensically Realistic Intercepted Telephone Speech Database (NFI-FRITS) is originated from speech of telephone recordings intercepted by Dutch law officials. The law officials laid down some conditions to use this material and one of the conditions is that the finalized database cannot disclose the identity of a particular speaker (Vloed, Bouten, & Van Leeuwen, 2014). The database contains real telephone speech from real police investigations. A similar speech database is AHUMADA 3 recorded by the Guardia Civil in Spain. The major problem in FSR is Database Mismatch and Database availability in context with forensics (Morrison, Rose, & Zhang, 2012; Ramos, Gonzalez-Rodriguez, Gonzalez-Dominguez, & Lucena-Molina, 2008).

The mismatch between training and testing conditions has become prior concern of researchers in the recent past. Various techniques have been developed to address this problem like speaker model synthesis (Teunen, Shahshahani, & Heck, 2000), several normalization techniques (Auckenthaler, Carey, & Lloyd-Thomas, Score normalization for text-independent speaker verification systems, 2000), factor analysis (Yin, Rose, & Kenny, 2007), feature mapping (Adami, Mihaescu, Reynolds, & Godfrey, 2003) and nuisance attribute projection (Solomonoff, Campbell, & Boardman, 2005). Many out of these techniques require parallel condition data, which is not contained by most of the publicly available speaker recognition databases (Haris B. C., et al., 2011). For robust speaker recognition the parallel condition data must be available as the surrounding conditions also affect the quality of the speech signal. For example, quality of speech sample recorded in a sound proof room will be much better than the speech sample recorded in a classroom or library. Sometimes noise in the speech samples is also desirable factor for designing a robust speaker recognition system, so that, the system can recognize the speaker even in the extreme conditions (Haris & Sinha, 2011). For this purpose a database should have the speech samples recorded in the different environments like a sound proof room, noisy class room, library, auditorium and market etc.

On the basis of literature review, we came out with a development framework of speaker recognition, in which we identified three main phases. In the first phase different factors that should be considered for the development of a robust speaker database are explained which includes session variability, channel variability, environmental noise, spoofing, whispering, variability in twins, age variability, physical and mental health of the speaker, language variability and regional\dialectal variability with-in the same language, intra-speaker and inter-speaker variability, variability due to style and situational mismatch and scarce availability of the databases considering most of the above said factors. As identified, the second phase of speaker recognition system is feature extraction which involves the selection of best feature for robust speaker recognition. Different type of features used for speaker recognition are discussed in detail[7], for example, short-term spectral features like MFCC, voice-source features like residual phase and glottal flow, spectro temporal features like modulation frequency, prosodic features like fundamental frequency, and high level features like idiolect etc [8].

III. PROPOSED SPEAKER RECOGNITION ALGORITHM

A small speaker database has been developed, as explained above, in which twenty speech samples are collected from each of the 100 different speakers including fifty male speakers and fifty female speakers. MFCCs are extracted for all the speakers. Then, ANN for a particular speaker is trained and this process is repeated for all the speakers one by one to get feature matrix of same size and then same is done with the Fuzzy Logic Technique. Fuzzy logic technique and ANN are explained in detail in the preceding chapter. The block diagram for the ANN\Fuzzy Logic based speaker identification system is shown in Figure 1.

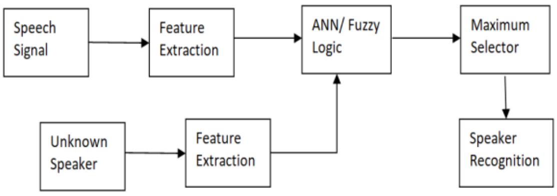

Fig 1. Proposed ANN\Fuzzy Logic based speaker identification system

The steps of the proposed algorithm are as follows:

Step-1: Database Collection: The first phase of any speaker recognition system is database collection. Twenty speech samples are collected from each of the 100 different speakers including fifty male speakers and fifty female speakers. These speakers belong to different regions of India i.e. they have different native language. Speakers are made to speak in three different languages namely, English, Hindi and native language of the speaker. To minimize channel mismatch problem, speech samples are recorded from five different sensors and three different styles as explained in the previous section.

Step-2: Pre-processing of Stored Speech Samples: After recording, the processing of the recorded speech samples is done. This pre-processing of recorded speech samples involves removal of noise, removal of silence, windowing and framing of the speech signals.

Step-3: Extraction of Mel-Frequency Cepstral Coefficients (MFCCs): MFCCs are extracted using MATLAB for all the available speech samples of all forty speakers according to the procedure detailed in the previous chapter.

Step-4: Training the ANNs: In this step, ANN having one hidden layer and five neurons is trained for each speaker. For training the ANN for first speaker, output parameter corresponding to the feature set of first speaker is set to 'one' and output parameters corresponding to all speakers except first speaker are set to 'zero'. This process adjusts the weights of this particular ANN accordingly. Now, this process is repeated for second speaker setting output parameter for second speaker equal to 'one' and for all speakers except second speaker are set equal to 'zero'. This will adjust the weights of this ANN according to second speaker. This process is repeated for each and every speaker. The whole procedure is repeated for different ANNs; one having two hidden layer and ten neurons; another ANN having three hidden layer and thirty neurons.

Step-5: Speech Signal of Unknown Speaker: Speech sample of the unknown speaker, which is obviously one of the forty speakers whose database has already been stored, is recorded and processed for silence and noise removal and windowing and framing is done. This speech signal of unknown speaker is called as test signal.

Step-6: Extraction of MFCCs from Test Signal: MFCCs from the processed speech sample of the unknown speaker are extracted repeating the same procedure followed in Step-3.

Step-7: Using these MFCCs as Input: The MFCCs extracted from speech signal of the unknown speaker are then applied as input to each of the trained ANN's corresponding to individual speakers And this process is repeated with the three different ANNs.

Step-8: Collecting the Output: The output from each of the ANN is collected and analyzed and the ANN which gives the maximum output corresponds to the unknown speaker and that speaker is declared as identified speaker.

Step-9: Using Fuzzy Logic: The whole procedure is repeated with Fuzzy Logic technique by applying the same set of features extracted from speech samples and these features are matched with the already stored features of known speaker using fuzzy logic and unknown speakers are recognized.

IV. EXPERIMENTATION AND RESULTS

Speech samples from the collected database are used for testing the model. The test speech samples are different than the training speech samples which are used to train the system for adjustment of the weights for each individual speaker. 100 speech samples are used for testing the efficiency of the proposed model. The results obtained with different systems are presents in the Table 1.

Table 1. Accuracy Rates

Classifier	No. of Neurons	No. of Hidden Layers	Accuracy Rate
ANN	5	1	68
	10	2	78
	15	3	74
Fuzzy Logic			72

MFCCs are extracted for all the speakers and these coefficients are used to train ANN\FL. The test speech samples, different than those of used for training, are used to evaluate the efficiency of ANNs and Fuzzy logic based speaker recognition systems using MATLAB. The results obtained are quite promising. The artificial neural network having five neurons and one hidden layer has identified thirty-four speaker correctly out of 100 test samples, thus, giving a success rate of 68% percent. Then, the MFCCs extracted from test samples are applied at second ANN having ten neurons and two hidden layers. This ANN recognized 96 speakers accurately providing an efficiency of 78%. The third ANN is consisting of thirty neurons and three hidden layers. On applying previously extracted MFCCs on this ANN, it gives an accuracy of seventy-four percent by recognizing thirty-seven speakers correctly out of fifty. The speaker recognition system using fuzzy logic has also been tested with the same set of test samples, which provided an efficiency of seventy-two percent by recognizing thirty-six speakers accurately. From the result obtained, it can be clearly seen that first ANN has given least accuracy and that is quite obvious because this system have simple architecture and have very small number of elements. On the other hand, it shows that increasing the number of elements and hence complexity in an ANN does not guarantee the better result as it is observed in case of third ANN having most complex architecture but still provided an accuracy lesser than the second ANN having somewhat simpler structure. Fuzzy logic also provided good accuracy but lesser than the second ANN. It proposes that ANN can give results better than fuzzy logic based systems. The obvious reason is that the samples recorded in a controlled and clean environment. Speaker recognition efficiency in uncertain environments can be enhanced using fuzzy logic technique as fuzzy logic is highly capable of analyzing uncontrolled and unidentified probabilistic signals.

V. CONCLUSION AND FUTURE SCOPE

Speaker recognition has utilization both in biometric as well as forensic applications. The proposed paradigm is used in biometric speaker recognition only. Further research can be done for the use of speaker recognition in the forensics field to supplement well-known method of DNA sampling for identification of any person and to produce it as evidence in the court of law. The features used in the current research are either MFCC or GFCC, but still there is scope for new techniques to be discovered and implemented to enhance the performance of the speaker recognition systems for better optimization of the results. For example, the combination of these two features or with some other feature can be used to obtain better performance.

The artificial neural network and fuzzy logic of type I techniques are utilized in this thesis for speaker recognition. Fuzzy logic of type II such as choquet fuzzy integral etc. can also be alternate for these techniques. Gaussian mixture model with universal background model can also be used. The literature review conducted in the thesis cannot claim to be exhaustive due to extensive nature of the domain and vast research work has been done by researchers in the field. Still there is scope for extending the literature review to bring out more granular details for reference to researchers. The speech samples in the database used are from forty speakers only. The experiment can be performed with more number of speakers from diverse domains as larger is the size of database better will be the efficiency of the system. Database used have addressed only noise and channel mismatch problems but still there are several problem remaining like whispered speech samples, spoofing, session variability and many more. To counter these problems, a robust database can be developed by collecting speech samples addressing this variability.

REFERENCES

1. Chorng-Shiuh Koong, Tzu-I Yang, and Chien-Chao Tseng, " A User Authentication Scheme Using Physiological and Behavioral Biometrics for Multitouch Devices", The Scientific World Journal, Volume 2014, Article ID 781234, 12 pages.
2. Kalyani CH," Various Biometric Authentication Techniques: A Review", Journal of Biometrics & Biostatistics, 2017, Vol 8(5): 371
3. Unichi Yamagishi ; Tomi H. Kinnunen ; Nicholas Evans ; Phillip De Leon ; Isabel Trancoso, "Introduction to the Issue on Spoofing and Countermeasures for Automatic Speaker Verification", IEEE Journal of Selected Topics in Signal Processing, Volume: 11, Issue: 4, June 2017.
4. George Saon, Gakuto Kurata," Computation and Language English Conversational Telephone Speech Recognition by Humans and Machines", Computation and Language, 6 Mar 2017
5. H. Soltau, G. Saon, and T. N. Sainath, "Joint training of convolutional and non-convolutional neural networks," to Proc. ICASSP, 2014
6. G. Saon, H.-K. Kuo, S. Rennie, and M. Picheny, "The IBM 2015 English conversational speech recognition system," in Sixteenth Annual Conference of the International Speech Communication Association, 2015
7. N Murali Krishna, J Sirisha Devi," A Novel Approach for Effective Emotion Recognition Using Double Truncated Gaussian Mixture Model and EEG", `I.J. Intelligent Systems and Applications, 2017, 6, 33-42
8. J Sirisha Devi, Dr. Srinivas Yarramalle, Siva Prasad Nandyala, "Speaker Emotion Recognition Based on Speech Features and Classification Techniques", I.J. Computer Network and Information Security, 2014, 7, 61-77

AUTHOR'S PROFILE

Dr. J Sirisha Devi was awarded B. Tech. in Computer Science and Engineering from Acharya Nagarjuna University -2003. She was awarded M. Tech. in Computer Science and Engineering from GITAM University, Visakhapatnam - 2010. She was awarded doctorate in the year 2016. Her research interests include Human Computer Interaction and Natural Language Processing.