ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2013

A Novel Approach for Text-Independent Speaker Identification Using Artificial Neural Network

Md. Monirul Islam[1], FahimHasan Khan[2], AbulAhsan Md. Mahmudul Haque[3]

Senior Software Engineer, Samsung Bangladesh R&D Center Ltd, Bangladesh [1]

Lecturer, Dept of CSE, Military Institute of Science and Technology, Dhaka, Bangladesh [2]

Assistant Professor, Dept of CSE, Rajshahi University of Engineering & Technology, Bangladesh [3]

ABSTRACT:This article presents the implementation of Text Independent Speaker Identification system. It involves two parts- "Speech Signal Processing" and "Artificial Neural Network". The speech signal processing uses Mel Frequency Cepstral Coefficients (MFCC) acquisition algorithm that extracts features from the speech signal, which are actually the vectors of coefficients. The backpropagation algorithm of the artificial neural network stores the extracted features on a database and then identify speaker based on the information. The direct speech does not work for the identification of the voice or speaker. Since the speech signal is not always periodic and only half of the frames are voiced, it is not a good practice to work with the half voiced and half unvoiced frames. So the speech must be preprocessed to successfully identify a voice or speaker. The major goal of this work is to derive a set of features that improves the accuracy of the text independent speaker identification system.

Keywords: Speaker identification, Text-Independent, Mel-Frequency Cepstral Coefficient, vector quantization, back-propagation.

I. INTRODUCTION

Human always identify speakers while they are talking to one another. The speakers may present in the same place or in different places. In this way a blind person can identify a speaker based solely on his/her vocal characteristics. Animals also use these characteristics to identify their familiar one [14]. Speaker identification is an important subject for research. Its development has come to a stage where it has been actively and successfully applied in a lot of industrial and consumer applications. It is being applied in biometrical identification, security related areas like voice dialling, banking by telephone, telephone shopping, database access services, information services, voice mail, passwords or keys, and remote access to computers.

A. *Identification Taxonomy*

Speaker recognition is usually divided into two different branches, speaker verification and speaker identification. Speaker identification can be further divided into two branches, Open-set speaker identification (Speakers from outside the training set may be examined) and closed-set speaker identification (The speaker is always one of a closed set used for training). Depending on the algorithm used for the identification, this task can also be divided into text-dependent (The speaker must utter one of a closed set of words) and text-independent identification (The speaker may utter any type of words). The identification taxonomy [11] is represented in Fig. 1.

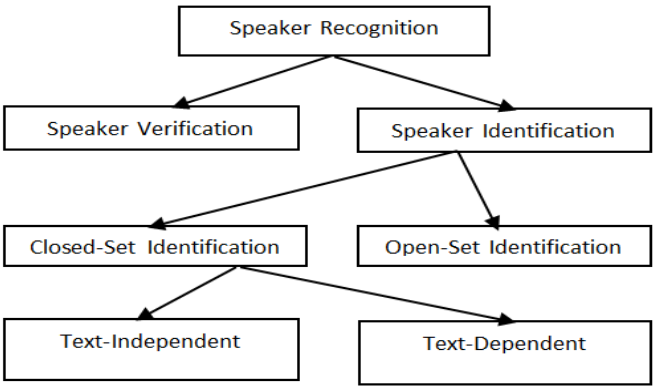

Fig.1 Identification Taxonomy

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2013

II. PHASES OF SPEAKER IDENTIFICATION SYSTEM

The process of speaker identification is divided into two main phases. The first phase is called the enrollment phase or learning phase. In this phase speech samples are collected from the speakers, and they are used to train their models. The collection of enrolled models is also called a speaker database. The processes of the enrollment phase [1] are represented in Fig. 2.

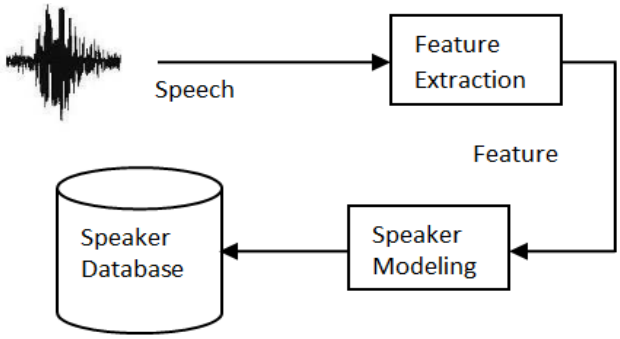

Fig.2Enrollment Phase

The second phase is called the identification phase, in which a test sample from an unknown speaker is compared against the speaker database. Both phases include the same initial step, feature extraction, which is used to extract speaker dependent characteristics from speech. The main purpose of the features extraction step is to reduce the amount of test data while retaining speaker discriminative information. The processes of the identification phase [1] are represented in Fig. 3.

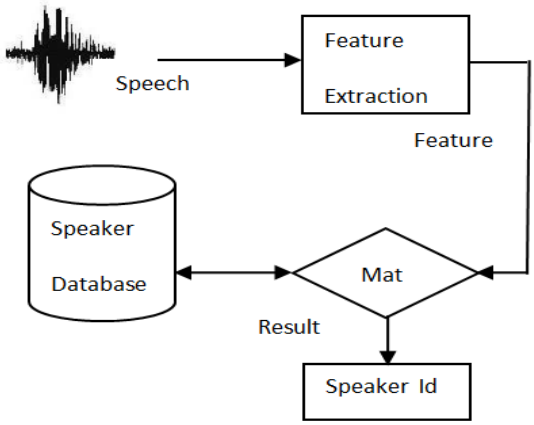

Fig.3 Identification Phase

However, these two phases are closely related, and so the identification algorithm usually depends on the modelling algorithm used in the enrollment phase.

III. THE SPEECH SIGNAL

A signal is defined as any physical quantity that varies with time, space, or any other independent variable or variables. Speech signals are examples of information bearing signal that evolve as functions of a single independent variable namely, time [24]. A speech signal is a complex signal, can be represented as

$$s(n) = h(n) * u(n)$$

Where, the speech signal s(n) is the convolution of a filter h(n) and some signal u(n). h(n) is also called the impulse response of the system. In our system (human body) h(n) is related with teeth, nasal cavity, lips etc. u(n) is approximately a periodic impulse train referred to as the pitch of speech, where pitch is synonymous with frequency.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2013

IV. FEATURE EXTRACTION

We can think about speech signal as a sequence of features that characterize both the speaker as well as the speech. The amount of data, generated during the speech production, is quite large while the essential characteristics of the speech process change relatively slowly and therefore, they require less data. Hence we can say that feature extraction is a process of reducing data while retaining speaker discriminative information [11]. A variety of choices for this task can be applied. Some commonly used methods for speaker identification is linear prediction and mel-cepstrum [4].

A. Mel-Frequency Cepstral Coefficients (MFCC's)

The cepstrum coefficients are the result of a cosine transformation of the real logarithm of the short time energy spectrum expressed on a Mel-frequency scale. This is a more robust, reliable feature set for speech recognition than the LPC coefficients. The sensitivity of the low order cepstrum coefficients to overall spectral slope, and the sensitivity of the high-order cepstrum coefficients to noise, has made it a standard technique. It weights the cepstrum coefficients by a tapered window so as to minimize these sensitivities, frame and these are used as the feature vector. In MFCC's, the main advantage is that it uses mel frequency scaling which is very approximate to the human auditory system. The coefficients generated by algorithm are fine representation of signal spectra with great data compression [11]. The process of extracting MFCC's from continuous speech is illustrated in Fig. 4.

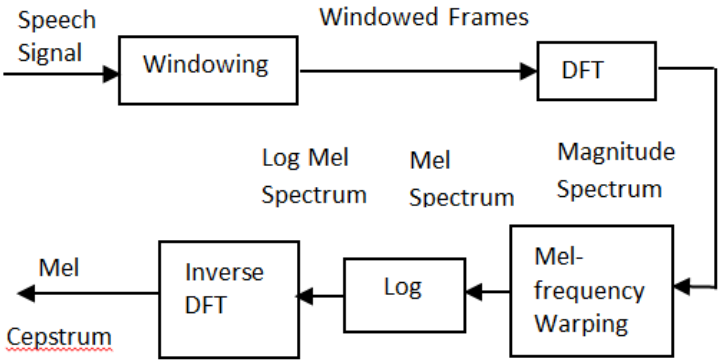

Fig. 4 The Steps in Creation of Mel-Cepstrum

V. DESIGN AND IMPLEMENTATION

The design and implementation of the text independent speaker identification system can be subdivided into two main parts: Speech signal processing and artificial neural network. The Speech signal processing contains speech signal acquisition and feature extraction. The neural network part consists of two main subparts: Learning and Identification.

The first part of text independent speaker identification i.e. speech signal processing consists of several steps. Firstly, we collected the speech signal using microphone and used low pass filter to remove noise from the signal. Then we detected the start and end point of the signal using the energy theory. Finally, we extracted the exact and effective features applying the feature extraction procedure. These extracted features were then fed into the neural network.

The second part of text independent speaker identification is the Artificial Neural Network (ANN). The first subpart of the artificial neural network is the learning and the second is the identification. For learning or training we applied the backpropagation learning algorithm. We adjusted the weight and threshold in learning phase, and saved into the database. In the identification phase, we used the database from learning algorithm to match the unknown speech signals.

A. Speech Signal Processing

We used MATLAB wavrecord function to record speech from real world. We used a microphone (MIC) as audio capture device. The speech signal is recorded at a sampling rate of 16 kHz and sampling length of 16 bits/sample.

1) Filtering: We used a low pass filter to remove noise from the speech signals. Here we considered the electrical noise that mixed with our input speech. Since the electrical signal has higher frequency than the speech signal, we filtered out the high frequency noise using a low pass filter. The cutoff frequency of the filter is 8 kHz.

2) Start and End Point Detection: We removed all non-speech samples and detected the start and end point of the recorded, temporal signals. We implemented this using an energy detection algorithm developed in a heuristic manner

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2013

from our data. Since none of our recordings contained speech in the first 100 ms of recording time, we analysed this time frame and generated an estimate of the noise floor for the speaking environment. Then we analysed each 20 ms frame and removed those frames with energy less than the noise floor.

3) Framing and Windowing: We framed the original vector of sampled values into overlapping blocks. Each block contains 256 samples with adjacent frames being separated by 128 samples. This yield a minimum of 50% overlaps to ensure that all sampled values are accounted for within at least two blocks. Since speech signals are quasi-stationary between 5 msec and 100 msec, 256 was chosen so that each block is 32 msec. 256 was chosen since it is a power of 2. This allows the use of the Fast Fourier Transform in subsequent stages. Here we used hamming window. The transform function of the window is

$$u(n) = x(n) * \mathrm{w}(n)\,, \qquad 0 \leq \mathrm{n} \leq \mathrm{N\text{-}1}$$
$$w(n) = 0.54 - 0.46\ cos\left(\frac{2\pi k}{K-1}\right), \qquad 0 \leq \mathrm{n} \leq \mathrm{N\text{-}1}$$

Where, N is the total No. of samples in a frame. Here N=256. The multiplicative scaling factor ensures appropriate overall signal amplitude.

4) *MFFC Generation and Vector Quantization:* We generated 12 mel frequency cepstral coefficients per frame and used as the feature vector. Before feeding this large number of features in the learning stage, we applied Vector Quantization technique on these features which compressed the large number of data sets into a smaller number of data. These data act as a representative of those data sets. For this work we took 60 centroids out of the MFCC coefficients and that concluded the feature extraction phase.

B. Learning

We trained the network on the individual patterns with a low error tolerance at the very beginning of the learning phase. The network then easily learned all the patterns within a short time. Because, a neural network needs relatively shorter time to learn the patterns with higher error tolerance values. In the next step, the same patterns were again trained to the neural network, but with a reduced error tolerance value. Since, the network already learnt all those patterns with a slightly higher error-tolerance, it never takes a long time to any extent for learning those patterns together with slightly less error tolerance. In this way, we conducted the error tolerance reduction and trained the neural network again and again. The network learnt all the patterns with a low error tolerance, and within a reasonably short time. A very simple multilayer Artificial Neural Network with a hidden layer is shown in Fig. 5 for better understanding of the parameters mentioned in the following section.

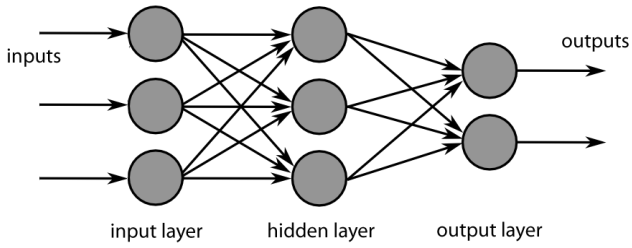

Fig. 5 Model of a Typical Multilayer Artificial Neural Network

For both the learning and the testing stage the network have the following parameters
No of neurons in the input layer: 60
No of neurons in the hidden layer: 20
No of neurons in the output layer: 5
Momentum coefficient = 0.5
Error rate= 0.005

C. Identification

After completing the learning phase we saved the weights. In the identification phase we fed the features from the speech signal that was to be identified into the network without having any target output. The network found the closest

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2013

matching output using the weights and thresholds stored before, and provided the corresponding speaker‘s index or id as the identified speaker.

VI. RESULTS AND PERFORMANCE ANALYSIS

The design and implementation of the text independent speaker identification system can be subdivided into two main parts: Speech signal processing and Artificial Neural Network.

A. Results

We performed our experiment considering different issues. We took different error rate (5 times) and completed the execution of the experiment again and again. In this case, we noticed how the error rate affects the identification of the speakers. We considered the identification capability of the network against the static speech signals and the instant speech recorded signal. The speaker database consisted of 16 speech samples from 8 speakers. Speakers were asked to read a given text in normal speed, under normal laboratory conditions. The same microphone was used for all recordings. For each speaker, two files were recorded; one for training and one for testing. Training and testing samples were recorded about 2 seconds long. The number of tests in each was 16. The experimented results are shown in the Table I.

TABLE I
RESULTS WITH VARYING ERROR RATE

Error Rate	**Successfully Identify**	**Error Result Shown**
0.1	4	12
0.05	6	10
0.03	8	8
0.01	12	4
0.005	14	2

We noticed that, our system works better at the lower error rate. The lower the error rate the higher the performance of the system. In our experiment for static speech signals we used eight speakers for the learning purpose. The experimented results for static voice aresummarized in Table II.

TABLE III
RESULTS FOR STATIC SPEECH SIGNALS

Total learned speaker	8
No of test with different input voice	16
Successfully identified speaker	14
Error result shown	2

We also used the same setup with eight speakers for the learning purpose for real time speech signals. The experimented results for real time voice areshown in Table III. We can see that the number of successfully identified speaker reduced in case of real time speech signals.

TABLE IIIII
RESULTS FOR REAL TIME SPEECH SIGNALS

Total learned speaker	8
No of test with different input voice	16
Successfully identified speaker	12
Error result shown	4

B. Performance

We measured the accuracy of this system as the success rate for the Identification of the speakers. It was measured using the following equation-

Success (%) = (Number of Success / Number of test) X 100 %.

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2013

We have shown the performance of our system with varying error rate in Table IV followed by graph in Fig. 6 showing the error rate and percentage of success. It can be easily observed from both the table and corresponding graph that the rate of success increased with less error rate.

TABLE IVV
SYSTEM PERFORMANCE WITH VARYING ERROR RATE

Error Rate	0.1	0.05	0.03	.01	0.005
% of Success	25.00	37.50	50.00	75.00	87.50

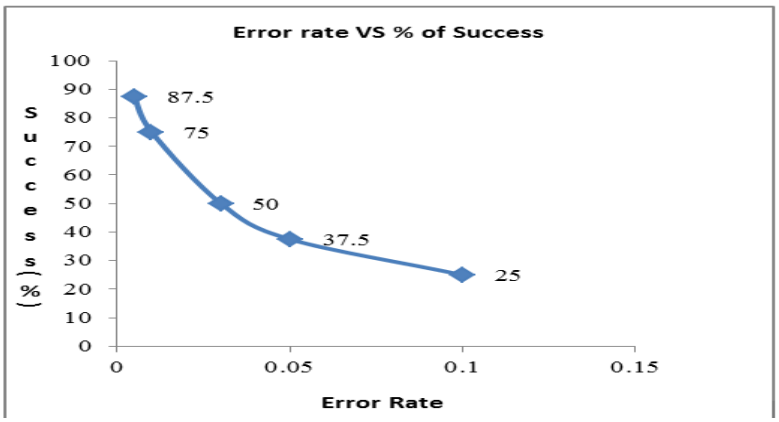

Fig. 6 System Performance with varying Error Rate

We have summarized the performances of our system with Static and Real time Speech Signal in the Table V and corresponding chart in Fig. 7. We can clearly observe that our system works better for the static speech signal than the real time (instantly recorded) speech signal. Our system was designed with a small number of input, output and hidden nodes. By increasing the number of input, output and hidden nodes the performance of the system can be further improved for bothstatic speech signal and real time speech signal.

TABLE V
SYSTEM PERFORMANCE WITH STATIC AND REAL TIME SPEECH SIGNALS

Static speech signals	87.50%
Real time speech signal	75.00%

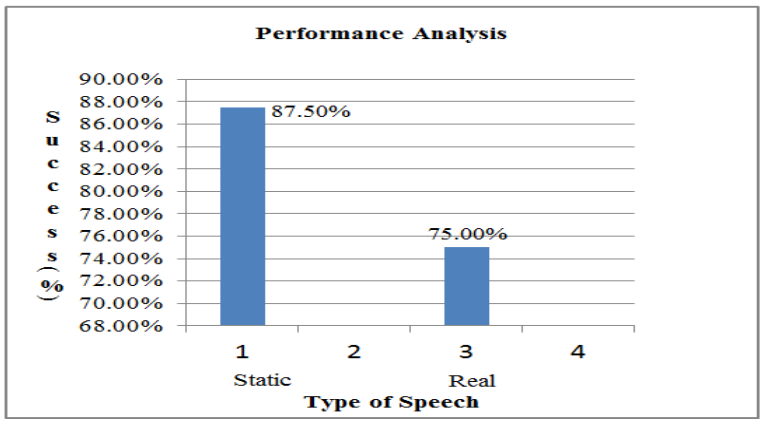

Fig. 7 System Performance with Static and Real time Speech Signal

VII. CONCLUSION

In this work, we studied and analyzed different techniques for speaker identification. We started from the identification background, which is based on the digital signal theory and modeling of the speaker vocal tract. Then we

ISSN (Print) : 2320 – 9798
ISSN (Online): 2320 – 9801

International Journal of Innovative Research in Computer and Communication Engineering
Vol. 1, Issue 4, June 2013

studied various techniques for reducing amount of test data or feature extraction techniques. Further, we studied most popular speaker modeling methods, which are commonly used in the speech recognition and speaker identification. From this work we see that in speaker identification process matching between test vectors and speaker models is the most time consuming part. It takes about 90 percent of all time spent on the identification. Therefore, optimization efforts should be concentrated on the matching optimization.

We concluded that, our system works better for the static speech signal than the real time speech signal. We can get more accuracy of the speaker identification system by using sufficient number of feature vectors as input and lowering the error rate to a certain level. The numbers of input, output and hidden nodes should also be increased. By using sufficient number of feature vectors and reducing error rate to a certain level then we can get near 100% accuracy in case of static speech signal and above 90% accuracy in case of real time speech signal. Based on our experiments and theoretical analysis, we can also conclude that our proposed speaker identification system is useful in practice.

REFERENCES

[1]. Kinnunen, T., and Li, H., "An overview of text-independent speaker recognition: From features to supervectors", Speech communication, Elsevier, Vol 52, Issue 1, pp.12–40, 2010.
[2]. Lu, X., & Dang, J., "An investigation of dependencies between frequency components and speaker characteristics for text-independent speaker identification", Speech communication, Elsevier, Vol 50, Issue 4, pp.312–322, 2008.
[3]. Gatica-Perez, D., Lathoud, G., Odobez, J.-M. andMcCowan, I., "Audiovisual probabilistic tracking of multiple speakers in meetings", IEEE Transactions on Speech and Audio Processing Vol 15-2,pp 601–616, 2007.
[4]. Niewiadomy, D., andPelikant, A.,"Digital Speech Signal Parametrization by Mel Frequency Cepstral Coefficients and Word Boundaries. Journal of Applied Computer Science", Vol15(2), pp71-81, 2007.
[5]. Bimbot, F., Bonastre, J. F., Fredouille, C., Gravier, G., Magrin-Chagnolleau, I., Meignier, S., ... and Reynolds, D. A., "A tutorial on text-independent speaker verification", EURASIP Journal on Applied Signal Processing, Vol 2004, pp 430-451, 2004.
[6]. Hasan, M. R., Jamil, M., andRahman, M. G. R. M. S.,"Speaker identification using Mel frequency cepstral coefficients", variations, Vol1, Issue 4, 2004.
[7]. Faundez-Zanuy, M. and Monte-Moreno, E., "State-of-the-art in speaker recognition", IEEE Aerospace and Electronic Systems Magazine, Vol 20-5, pp.7–12, 2005.
[8]. Linde, Y., Buzo, A., Gray and R.M., "An algorithm for vector quantizer design", IEEE Transactions on Communication,Vol 28, pp 84–95, 1980.
[9]. Becker, T., "MFCC Feature Vector Extraction Using STx", Acoustics Research Institute, Austrian Academy of Sciences, 2007.
[10]. Pawar, R.V., Kajave, P.P. and Mali, S.N., "Speaker Identification using neural Networks", World Academy of Science, Engineering and Technology, 2005.
[11]. Karpov, E., "Real-Time Speaker Identification", Department of Computer Science, University of Joensuu, 2003.
[12]. Nilsson. M and Ejnarsson, M., "Speech Recognition using Hidden Markov Model performance evaluation in noisy environment", Department of Telecommunication and Signal Processing, Blekinge Institute of Technology, Ronneby, 2002.
[13]. Le, C.G.,"Application of a back propagation neural network to isolated-word speech recognition", Aval Postgraduate School Monterey Ca, 1993.
[14]. Love, B.J., Vining, J. and Sun, X., "Automatic Speech Recognition using Neural Networks",The University of Texas at Austin, 2004.
[15]. Proakis, J. G., and Manolakis, D. G., "Digital signal processing: principles, algorithms, and applications", Pearson Education India, 2001.

BIOGRAPHY

Md. Monirul Islam Completed BSc degree in Computer Science and Engineering from Rajshahi University of Engineering & Technology (RUET), Bangladesh, in 2009.He is currently working as aSenior Software Engineer in Samsung Bangladesh R&D Center Ltd, Bangladesh. His research interests include artificial neural networks, mobile and ubiquitous computing, Software Engineering.

FahimHasan Khan was awarded BSc degree in Computer Science and Engineering from Rajshahi University of Engineering & Technology (RUET), Bangladesh, in 2009. He is currently working as a Lecturer in Department of Computer Science and Engineering in Military Institute of Science and Technology (MIST), Bangladesh and also as an MSc student at the Department of Computer Science and Engineering in Bangladesh University of Engineering & Technology (BUET). His research area includehigh dimensional databases, artificial neural networks, computer graphics and image processing.

AbulAhsan Md Mahmudul Haque received BS degree in Computer Science and Information Technology from Islamic University of Technology (IUT), Bangladesh, in 2002, and MS degree in Security and Mobile Computing from the Helsinki University of Technology (HUT), Finland, in 2008. He is currently a PhD student and research fellow at the Department of Computer Science, University of Tromsø, Norway. His research interests include peer-to-peer networking, mobile computing, workflow, Web services and services orchestration.