ISSN:2319-6890)(online),2347-5013(print)
01 Dec. 2014

Text Dependent Speaker Identification Using a Bayesian network and Mel Frequency Cepstrum Coefficient

[1]Mohd. Manjur Alam , [2] Md. Salah Uddin Chowdury ,[3] Niaz Uddin Mahmud , [4]Shamsun Nahar Shoma , [5]Md. Abdul Wahab

Dept. of CSE, BGC Trust University Bangladesh

manjur66@gmail.com, Schowdhury_cse@yahoo.com, niazumahmud@gmail.com, shomathetics@yahoo.com , wahab_nstu@yahoo.com

***Abstract:** Speaker identification is a biometric technique. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. Speaker Recognition technology has recently been used in large number of commercial areas successfully such as in voice based biometrics; voice controlled appliances, security control for confidential information, remote access to computers and many more interesting areas. A speaker identification system has two phases which are the training phase and the testing phase. Feature extraction is the first step for each phase in speaker recognition. Many algorithms are suggested by the researchers for feature extraction. In this work, the Mel Frequency Cepstrum Coefficient (MFCC) feature has been used for designing a text dependent speaker identification system. While, in the identification phase, the existing reference templates are compared with the unknown voice input. In this thesis, a Bayesian network is used as the training/recognition algorithm which makes the final decision about the specification of the speaker by comparing unknown features to all models in the database and selecting the best matching model. i, e. the highest scored model. The speaker who obtains the highest score is selected as the target speaker.*

Keywords- **Mel Frequency Cepstrum Coefficient (MFCC), Bayesian network(BN), Speaker Identification (SI). graphical models (GMs), directed a cyclic graph(DAG),Joint Probability Distribution (JPD),Discrete Fourier Transform(DFT).**

I. INTRODUCTION

Speaker identification system is classified into text dependent and text independent recognition systems [1]. In text dependent recognition systems, systems know the text spoken by person. Fixed or prompted phrases or texts are used by the speakers. Same text must be used for enrollment and verification. They generally are more accurate and improve system performance as system has knowledge of spoken text. In text independent recognition, system does not know the text to be spoken by person. Text entered during enrollment and test is different. It is more flexible but also more difficult. Recognizing natural speech is a challenging task as we are unable to see the speech signals. Amplitude, pitch, phonetic emphasis etc. are the various speech parameters that are used in speech recognition systems. This identification is text-dependant.

The human speech contains numerous discriminative features that can be used to identify speakers [2]. Speech contains significant energy from zero frequency up to around 5 kHz. The objective of automatic speaker recognition is to extract, characterize and recognize the information about speaker identity. Anatomical structure of the vocal tract is unique for every person and hence the voice information available in the speech signal can be used to identify the speaker. Recognizing a person by her/his voice is known as speaker recognition. Since differences in the anatomical structure are an intrinsic property of the speaker, voice comes under the category of biometric identity. Using voice for identity has several advantages. One of the major advantages is remote person authentication. The property of speech signal changes markedly as a function of time. To study the spectral properties of speech signal the concept of time varying Fourier representation is used. However, the temporal properties of speech signal such, as energy, zero crossing, correlation etc are assumed constant over a short period. Its characteristics are short-time stationary .Therefore, using hamming window, Speech signal is divided into a number of blocks of short duration so that normal Fourier transform can be used. In this thesis the most important thing is to extract the feature from the speech signal. The speech feature extraction in a categorization problem is about reducing the dimensionality of the input-vector while maintaining the discriminating power of the signal. As we know from the above fundamental formation of speaker identification that the number of training and test vector needed for the classification problem grows exponential with the dimension of the given input vector, so we need feature extraction. But extracted feature should meet some criteria while dealing with the speech signal.

Such as:

- Easy to measure extracted Speech features.
- Distinguish between speakers while being lenient of intra speaker variability's.
- It should not be susceptible to mimicry.
- It should show little fluctuation from one speaking environment to another.
- It should be stable over time.
- It should occur frequently and naturally in speech.

In this work, the Mel frequency Cepstrum Coefficient (MFCC) feature has been used for designing a text dependent speaker identification system. A speaker identification system consists of two phases which is the training phase and the testing phase. In the training phase, the speaker voices are recorded and processed in order to generate the model to store in the database. While, in the identification phase, the existing reference templates are compared with the unknown voice input. The identification is done using BNs. The code is developed in the MATLAB environment and performs the identification satisfactorily.

II. PHASES OF SPEAKER IDENTIFICATION

Like any other pattern recognition systems, speaker recognition systems also involve two phases namely, training and testing [2]. Training is the process of familiarizing the system with the voice characteristics of the speakers registering. Testing is the actual recognition task. We make the final decision about the identity of the speaker by comparing unknown features to all models in the database and selecting the best matching model. i, e. the highest scored model. This is done using BN.

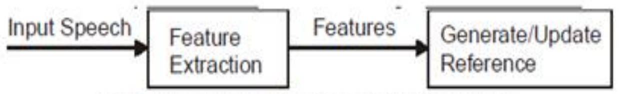

Fig.1. The block diagram of training phase.

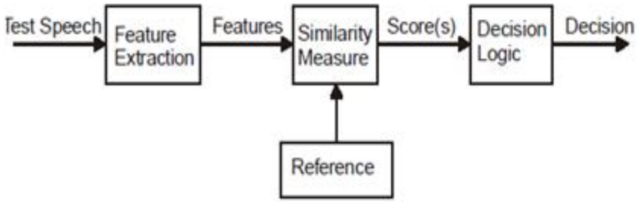

Fig.2.The block diagram of the testing phase

III. SPEECH FEATURE EXTRACTION

In this thesis the most important thing is to extract the feature from the speech signal. We are using the Mel Frequency Cepstral Coefficients (MFCC) technique to extract features from the speech signal and compare the unknown speaker with the exits speaker in the database. MFCCs are based on the known variation of the human ear’s critical bandwidths with frequency, filters spaced linearly at low frequencies and logarithmically at high frequencies have been used to capture the phonetically important characteristics of speech. This is expressed in the *mel-frequency* scale, which is linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000 Hz. Fig3. below shows the complete pipeline of Mel Frequency Cepstral Coefficients.

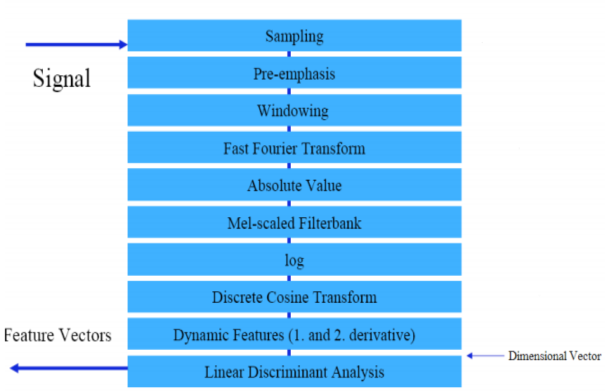

Fig.3. Mel Frequency Cepstral Coefficients

A. FRAMING AND WINDOWING

As shown in the figure (Fig3) the speech signal is slowly varying over time and it is called quasi stationery. When the speech signal is examined over a short period of time such as 5 to 100 milliseconds, the signal is reasonably stationery, and therefore these signals are examined in short time segment.

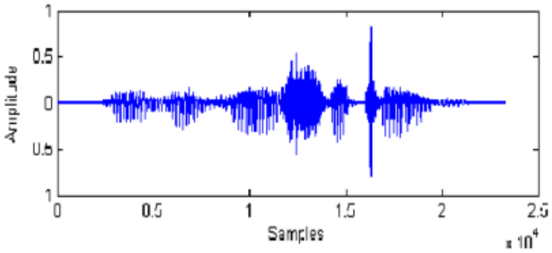

Fig.4. Framing and Windowing

B. HAMMING WINDOW

Hamming window is also called the raised cosine window. The equation and plot for the Hamming window shown in Fig.4.

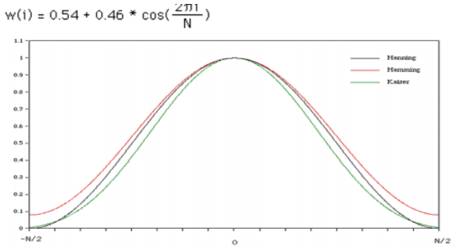

Fig.4. Hamming Window

C. FAST FOURIER TRANSFORM

FFT is a very important mathematical tool.FFT algorithms are, the discrete time sequence *x(n)* is transformed into corresponding discrete frequency sequence *X[k]*. The FFT is a fast algorithm to implement the Discrete Fourier Transform (DFT), which is defined on the set of *N*, samples {*xn*}, as follow[3]:

$$X_k = \sum_{n=0}^{N-1} x_n \, e^{-j2\pi kn/N} \qquad k = 0,1,2,\ldots N \qquad (1)$$

When X_k’s are generally complex numbers and we only consider their absolute frequencies here. The resulting ssequence $\{X_k\}$ is interpreted as follow: positive frequencies $0 \leq f < F_s/2$ correspond to values $0 \leq n \leq N/2-1$, while negative frequencies $-F_s/2 < f < 0$ correspond to $N/2+1 \leq n \leq N-1$. Here, F_s denote the sampling frequency.

D. MEL FREQUENCY WARPING

For each sound with an actual frequency, *f* , measured in Hz, a subjective frequency is measured on a scale called the "Mel scale" Mel-frequency can be approximated by

$$\mathrm{Mel}(f)=1127\ln\left(\frac{f}{100}+1\right) \qquad (2)$$

Where *f* in Hz, is the actual frequency of the sound[3].

E. CEPSTRUM

It is defined as the inverse Fourier transform of the logarithm of the magnitude of the Fourier transform; i.e.

cepstrum= iffft(log(|fft(signal)|)) (3)

Where the function iff() returns the inverse discrete Fourier transform[3]

F. TRIANGULAR FILTERS BANK

We define a triangular filter-bank with *M* filters (*m*=1, 2,…,*M*) and *N* points Discrete Fourier Transform (DFT) (*k*=1,2,…,*N*), where H_m[k] is the magnitude (frequency response) of the filter given by[3]:

$$H_m(k) = \begin{cases} 0, & k<f[m-1] \\ \frac{(k-f[m-1])}{(f[m]-f[m-1])}, & f[m-1]\leq k\leq f[m] \\ \frac{(f[m+1]-k)}{f[m+1]-f[m]}, & f[m]\leq k\leq f[m+1] \\ 0, & k>f[m+1] \end{cases} \qquad 4$$

Such filters compute the average spectrum around each center frequency with increasing bandwidths, and they are displayed in Fig6.

the Let $f_l$and f_h be the lowest and highest frequencies of the filter-bank in Hz, Fs the sampling frequency in Hz, *M* the number of filters, and *N* the size of the Fast Fourier Transform. The boundary points are uniformly spaced in Mel-scale

$$(f[m] = \left(\frac{N}{F_s}\right) Mel^{-1} \left(Mel(f_l) + m\frac{Mel(f_h)-Mel(f_l)}{M}\right), \quad (5)$$
$$0 \leq m \leq M+1$$

Where Mel(*f*) is given by (1) and $Mel^{-1}(f_l)$ is its inverse given by (6)

$$Mel^{-1}(f) = 700(\exp\left(\frac{f}{1127}\right) - \qquad (6)$$

CALCULATION OF MFCCS:

Given the DFT of the input signal, x[n][3]

$$X_a[k] = \sum_{n=0}^{N-1} x[n]e^{-\frac{j2\pi nk}{N}}, \quad 0 \leq k \leq N \qquad (7)$$

In most implementations of speech recognition, a short-time Fourier analysis is done first. Then the values of DFT are weighted by triangular filters.

The result is called Mel-frequency power spectrum which is defined as:

$$S[m] = \sum_{k=1}^{N} X_a[K]^2 H_m[K], 0 < m \leq M \qquad (8)$$

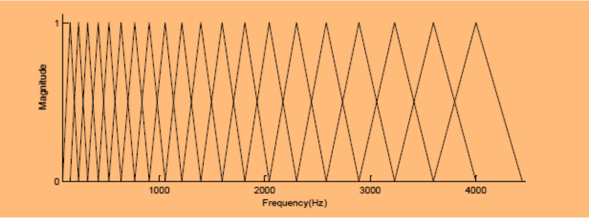

Fig 6. Filter bank for generating Mel-Frequency Cepstrum Coefficients.

Where X_a [k]2 is called power spectrum. Finally, a discrete cosine transform (DCT) of the logarithm of S[m] is computed to form the MFCCs as

$$mfcc[i] = \sum_{m=1}^{M} \log[s[m] \cos\left[i\left(m-\frac{1}{2}\right)\right]\frac{\pi}{M} \qquad (9)$$

i=1,2…..L

Where *L* is the number of cepstrum coefficients.

IV. BAYESIAN NETWORKS (BNS):

In this thesis, we use the Bayesian networks (BNs) method as the training/recognition algorithm [6]. It is also known as belief net-works(or Bayes nets for short),belong to the family of probabilistic graphical models (GMs). BNs correspond to another GM structure known as a directed a cyclic graph (DAG). A more formal definition of a BN can be given [7]; A Bayesian net-work B is an annotated acyclic graph that represents a JPD over a set of random variables V. The net-work is defined by a pair,B=(G,Ө) where G is the DAG whose nodes $X_{1,}$ $X_{2,}$ $X_{3,}\ldots\ldots X_N$ represents random variables, and whose edges represent the direct dependencies between these variables. The graph G encodes independence assumptions, by which each variable X_i is independent of its nondescendents given its parents in G. The second component, Ө denotes the set of parameters of the network. This set contains the parameter $\Theta_{x_i\ |\pi_i} = \quad P_B\ \ (X_i\ |\pi_i)$ for each realization, x_i of X_i conditioned on π_i the set of parents of X_i in G. Accordingly, B defines a unique JPD over V, namely:

$$P_B(X_{1,}\ X_{2,}\ X_{3,}\ldots\ldots X_N) = \prod_{i=1}^{n} P_B\ \ (X_i\ |\pi_i)\ \ = \prod_{i=1}^{n} \Theta_{x_i\ |\pi_i} \quad (10)$$

A. PROBABILITIES MEASUREMENT

Probabilities from states to observations are measured from matching MFCCs from states to observations. The matching MFCCs are proportional to probability. Here the mfcc of each frame of the training data is compared to that of the testing data with sequence for each utterance

B. SCORES

For each speaker model, probability score for the unknown observation sequence is computed through individual BN using eq.(10). The speaker whose model produces the highest probability score and matches the ID claimed is then selected as the client speaker.

V. EXPERIMENTAL RESULTS

I have generated MFCCs from five speakers saying “CSE” using Matlab and recorded for speaker recognition. Each speaker has uttered the same word for three times. Here the duration was 3.0 seconds. The speech was recorded using a sampling rate of 44100 Hz with 16 bits per sample. The waveform corresponding to the utterance is shown at fig7. Speech waveform sampled at 1600Hz.

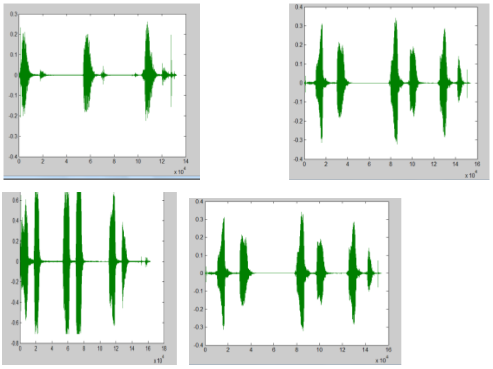

Fig.7: Wave Forms(Manjur, Parvez, Sagor, Kayes)

A Feature Extraction

We used a population of 4 speakers, with 3 utterances (same word) for each speaker. Each utterance was divided into 2210frames (total 6630). The average length of each frame was about 32 milliseconds (256 samples), Then MFCC were calculated for each frame using Matlab. Similarly the features are extracted from testing data. Here the mfcc of each frame of the training data is compared to the of the testing data for each utterance. I have selected only 150mfccs from mfccs for each speaker: 50 from first utterance, 50 from second one and 50 from third one in training and testing data.

B Matching Between Training & Testing Data

The mfcc of each frame of the training data is compared to that of the testing data for each utterance. The two mfccs which distance between them within 1.5 are similar. This is done from the both sides.

Matching MFCCS=Total MFCCs – numbers of uncommon MFCCS in training data & Testing data. ……(11)

	Unknown Speaker		
	1^{st} utterance	2nd utterance	3^{rd} utterance
Speaker	matching mfccs	matching mfccs	matching mfccs
Manjur	42	49	49
Parvez	45	1	49
Sagor	27	49	49
Kayes	13	49	49

Table1: Matching between training & testing data set for each speaker

C. Bayesian Network For Speaker1(Manjur):

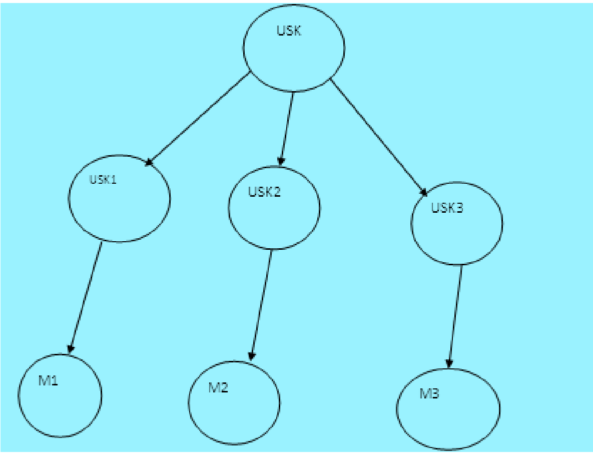

Fig 8: BN for **Speaker1(Manjur)**

Here
USK= average frame mfcc of unknown speaker,
USK1=1^{st} utterance of unknown speaker
USK1=2^{nd} utterance of unknown speaker
USK1=3^{rd} utterance of unknown speaker
M1=1^{st} utterance of Manjur
M2=2^{nd} utterance of Manjur
M3=3^{rd} utterance of Manjur

Probability measurement(Manjur):

P(USK,USK1,USK2,USK3,M1,M2,M3)
= P(USK)P(USK1|USK) P(USK2|USK) P(USK3|USK) P(USK1|M1) P(USK2|M2) P(USK3|M3)

Here,
P(USK)= average matching mfccs/ total mfccs =46.66/50= .933
P(USK1|USK)= 1/3 =.33
P(USK2|USK)= 1/3=.33
P(USK3|USK)= 1/3 =.33
P(USK1|M1)= matching mfccs between 1^{st} utterance of unknown speaker to 1^{st} one of Manjur/ total mfccs,
= 42/50 =.84
P(USK2|M2)= matching mfccs between 2^{nd} utterance of unknown speaker to 2^{nd} one of Manjur/ total mfccs
=49/50 =.98 ,
P(USK3|M3)= matching mfccs between 3^{rd} utterance of unknown speaker to 3^{rd} one of Manjur/ total mfccs
=49/50 =.98 ,

Hence,
P(USK,USK1,USK2,USK3,M1,M2,M3)= .933 * .33 * .33 * .33 * .84 * .98 * .98 =.027049

D. A Bayesian network for Speaker2(Parvez):

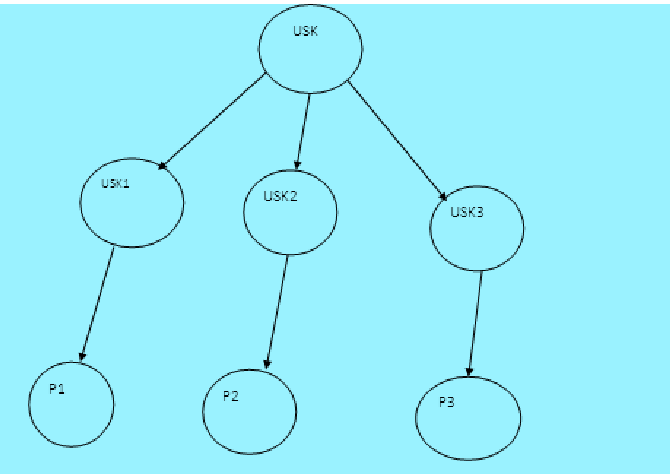

Fig 9: BN for **Speaker1(Parvez)**

Here
USK= average frame mfcc of unknown speaker,
USK1=1^{st} utterance of unknown speaker
USK1=2^{nd} utterance of unknown speaker
USK1=3^{rd} utterance of unknown speaker
P1=1^{st} utterance of Parvez
P2=2^{nd} utterance of Parvez
P3=3^{rd} utterance of Parvez

Probability measurement:

P(USK,USK1,USK2,USK3,P1,P2,P3)
= P(USK)P(USK1|USK) P(USK2|USK) P(USK3|USK) P(USK1|P1)P(USK2|P2) P(USK3|P3)

Here,
P(USK)= average matching mfccs/ total mfccs =31.66/50 = .6332
P(USK1|USK)= 1/3 =.33
P(USK2|USK) = 1/3 =.33
P(USK3|USK)= 1/3=.33
P(USK1|P1)= matching mfccs between 1^{st} utterance of unknown speaker to 1^{st} one of parvez/ total mfccs,
=45/50 = 0.9
P(USK2|P2)= matching mfccs between 2^{nd} utterance of unknown speaker to 2^{nd} one of parvez/ total mfccs
,=1/50 = 0.02

P(USK3|P3)= matching mfccs between 3rd utterance of unknown speaker to 3rd one of parvez/ total mfccs
= 49/50 = 0.98

Hence,
P(USK,USK1,USK2,USK3,P1,P2,P3)=.6332 * .33 * .33 * .33 * .9 * .02 * .98 =.000401

E.A Bayesian network for Speaker3(Sagor):

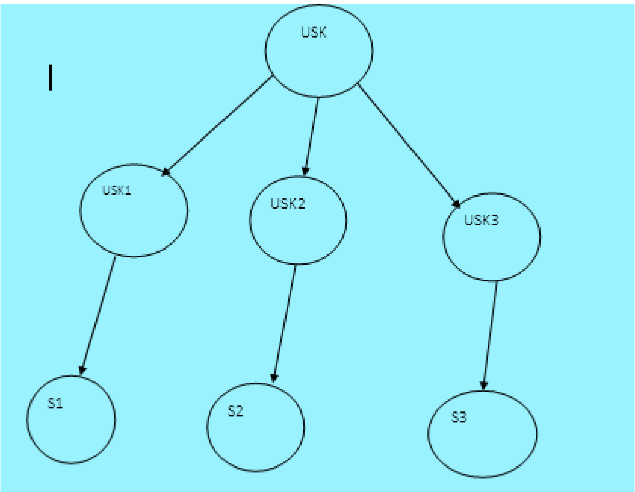

Fig 10: BN for **Speaker3(Sagor)**

Here
USK= average frame mfcc of unknown speaker,
USK1=1st utterance of unknown speaker
USK1=2nd utterance of unknown speaker
USK1=3rd utterance of unknown speaker
S1=1st utterance of Sagor
S2=2nd utterance of Sagor
S3=3rd utterance of Sagor

Probability measurement:

P(USK,USK1,USK2,USK3,S1,S2,S3)
= P(USK)P(USK1|USK) P(USK2|USK) P(USK3|USK) P(USK1|S1) P(USK2|S2) P(USK3|S3)

Here,
P(USK)= average matching mfccs/ total mfccs =41.66/50 =0.8332
P(USK1|USK)= 1/3 = 0.33,
P(USK2|USK)= 1/3 = 0.33,
P(USK3|USK)= 1/3 = 0.33,
P(USK1|S1)= matching mfccs between 1st utterance of unknown speaker to 1st one of Sagor/ total mfccs =27/50 = .54
P(USK2|S2)= matching mfccs between 2nd utterance of unknown speaker to 2nd one of Sagor/ total mfccs
=49/50 =0.98
P(USK3|S3)= matching mfccs between 3rd utterance of unknown speaker to3rd one of Sagor/ total mfccs
=49/50 =0.98

Hence,
P(USK,USK1,USK2,USK3,S1,S2,S3)=.833 * 0.33 * 0.33 * 0.33 * .54 * .98 *.98 = .0155

5.6A Bayesian network for Speaker4(Kayes):

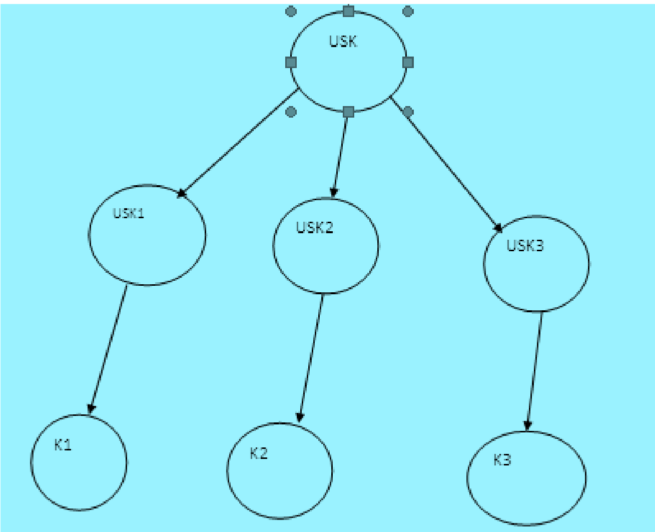

Fig 11: BN for Speaker4(Kayes)

Here
USK= average frame mfcc of unknown speaker,
USK1=1st utterance of unknown speaker,
USK1=2nd utterance of unknown speaker
USK1=3rd utterance of unknown speaker
S1=1st utterance of Kayes
S2=2nd utterance of kayes
S3=3rd utterance of Kayes

Probability measurement:

P(USK,USK1,USK2,USK3,K1,K2,K3)
=P(USK)P(USK1|USK)P(USK2|USK)P(USK3|USK)P(USK1|K1) P(USK2|K2) P(USK3|K3)
P(USK)= average matching mfccs/ total mfccs =37/50
= .74
P(USK1|USK)= 1/3 =.33
P(USK2|USK)= 1/3 =.33
P(USK3|USK = 1/3 = .33
P(USK1|K1)= matching mfccs between 1st utterance of unknown speaker to 1st one of Kayes/ total mfccs
= 13/50= .26
P(USK2|K2)= matching mfccs between 2nd utterance of unknown speaker to 2nd one of Kayes/ total mfccs
= 49/50 =.98
P(USK3|K3)= matching mfccs between 3rd utterance of unknown speaker to 3rd one of Kaye/ total mfccs
= 49/50=.98

Hence,
P(USK,USK1,USK2,USK3,K1,K2,K3)=.74 * .33 * .33 * .33 *.26 * .98 * .98 = .00664

VI. RESULT

The highest score is obtained from Manjur, hence the unknown speaker was Manjur.

VII. CONCLUSION

In our work we trained the system for each speaker for same word ‘CSE’ which the speaker utters for three times. We then specified that particular speaker by comparing each registered speaker in the training phase. In the training phase we actually build a reference model for a particular speaker and compared that stored reference model against the input speech by evaluating scores using BN and speaker identification with the best score is done. The system developed is speaker and text dependent and moderately tolerant to background noise hence it is a very efficient system. Due to dynamic nature of the speech signals, current speaker identification systems produce reasonable results , but still lack the necessary performance if they are to be used the general public. The variability in speech is mainly caused by the length of the vocal track, varying pitch and speaking rate as well as different accents and speaking styles This paper challenged this problem and the results obtained did not only improve the identification rate but also improved the reliability of the SI system. The system uses the MFCC feature extraction technique and BN for evaluating the scores which proved to be a great success. The experiment is carried out in noise free environment. The system correctly identified the speaker trained for a particular word by comparing the input speech for that word against the stored reference model for that word.

VIII.REFERENCES

i. *Speaker Verification Using Vector Quantization And Hidden Markov Model. Mohd Zaizu Ilyas, Member, IEEE, Salina Abdul Samad, Senior Member, IEEE, Aini Hussain, Member , IEEE And Khairul Anuar Ishak, Member, IEEE, "**The 5th Student Conference On Research And Development –Scored 2007 11-12 December 2007, Malaysia".***

ii. *MFCC And Its Applications In Speaker Recognition, Vibha Tiwari. Deptt. Of Electronics Engg., Gyan Ganga Institute Of Technology And Management, Bhopal, (MP) INDIA (Received 5 Nov., 2009, Accepted 10 Feb., 2010), "International Journal On Emerging Technologies **1**(1): 19-22(2010) **ISSN : 0975-8364".***

iii. *Text-Independent Speaker Identification Using Hidden Markov Model Sayed Jaafer Abdallah, Izzeldin Mohamed Osman, Mohamed Elhafiz Mustafa, College Of Computer Science And Information Technology, Sudan University Of Science And Technology, "World Of Computer Science And Information Technology Journal (WCSIT) ISSN: 2221-0741 Vol. 2, No. 6, 203-208, 2012 Khartoum, Sudan.".*

iv. *Speaker Identification Using Mel Frequency Cepstral Coefficients, Md. Rashidul Hasan, Mustafa Jamil, Md. Golam Rabbani Md. Saifur Rahman, Electrical And Electronic Engineering, Bangladesh University Of Engineering And Technology, "Dhaka-1000, **3**rd International Conference On Electrical & Computer Engineering ICECE 2004, 28-30 December 2004, Dhaka, Bangladesh".*

v. *Speaker Recognition Using MFCC Front End Analysis And VQ Modeling Technique For Hindi Words Using MATLAB, Nitisha M.Tech. (Pursuing) Hindu College Of Engineering Sonipat, Haryana India, Ahu Bansal Assistant Professor Hindu College Of Engineering Sonipat, Haryana, India, "International Journal Of Computer Applications (0975 – 8887) Volume 45– No.24, May 2012".*

vi. *Ben-Gal I., Bayesian Networks, In Ruggeri F., Faltin F. & Kenett R., Encyclopedia Of Statistics In Quality & Reliability, Wiley & Sons (2007).*

vii. *Friedman, N., Geiger, D. & Goldszmidt, M. (1997).Bayesian Network Classifiers, Machine Learning **29**,131–163.*