Voice Activated E-Learning System for the Visually Impaired

S.Asha
Department of Computer Science and Engineering
College of Engineering, Guindy,
Anna University, India

C.Chellappan
Department of Computer Science and Engineering
College of Engineering, Guindy,
Anna University, India

ABSTRACT
E-learning has become an important tool for learners to acquire information and knowledge. However visually impaired people have no or very little access to this tool, since interface suitable to them are unavailable. The Voice Activated E learning System can provide a solution to this problem. Developing this system is meant to assist visually impaired students in learning, a desired subject, from the system, in a convenient way using their voice commands. This system consists of two major subsystems; namely Speaker Verification and Speech Recognition subsystem. In the speaker verification subsystem, Mel-Frequency Cepstral Coefficients (MFCC) is used for Feature extraction and Vector Quantization (VQ) algorithm is used for codebook generation. In the speech recognition subsystem, MFCC and dynamic programming (DP) are used. Experimental results show an accuracy of 96% in speaker verification subsystem and 89% in speech recognition subsystem.

1. INTRODUCTION

The e-Learning market has been slow in developing courses for use with screen readers, speech recognition software and other adaptive technologies. The technology is there but people can't use it in a way that could certainly benefit them most. Although present software systems are often very sophisticated and user-friendly they are not very convenient for visually impaired people. The present development in human-computer interaction and spoken language dialog systems brings new hopes. Going one step further, *true usability with these technologies must be considered very early in the design process*. E-learning has become an important tool for learners to acquire information and knowledge. However visually impaired people have no or very little access to this tool since no interface suitable to them are available. According to WHO, approximately 340 million people in this world are visually impaired. So there is a great need to develop a user friendly easily navigable voice interface to the E-learning environment. The main objective of this work is to create an E learning system that can perform text-independent user verification and as well can respond to speech commands, which allows enhancement of the quality of the interaction between visually impaired user and the e-learning platform.

Once the user enrolls into the system, the user will be guided through the system, using audio instructions. The user can choose various options in the system, with the help of speech inputs and few keyboard operations.

1.1 E-Learning Environment

Every person having physical impairments has an equal right to obtain education. This project presents an alternate use of E-Learning technology to support visually impaired students. The major problem for visually impaired is to access the resources being used in E Learning systems. These people mostly use screen reader software (JAWS, Window-Eyes etc) because Braille is expensive and slow [8].

Although the present software systems are often sophisticated and user-friendly, they are usually not very convenient for the visually impaired people. The reason is the graphical interface and absence of the features fulfilling special needs of the visually impaired. Speech synthesizer and screen reader software still represent basic functionalities that are used by the visually impaired to obtain information by means of a computer [21].

Since visual access interfaces are of no use to visually impaired and are problematic for those with vision disabilities, the need to create other special access interfaces arose, which would support the information flow through alternative sensory routes such as hearing and touch [2].

Basically, today's elearning systems faces two major issues: i) Audio output using Screen Reader Software and ii) User commands only through keys and not Voice Input. Firstly, most of the visually impaired use "JAWS" screens reader software for operating computer. Its accent is American and difficult to understand by the Asians. According to a study conducted by "Enabling Dimensions' in India, "only 3 out of 10 students are able to overcome the difficulty of accent" [8] and also Screen Readers just read out the contents of the screen aloud, but does not allow interactions with the user[8]. Secondly, the navigation through the pages is realized by pressing the keys according to the instigation of the corresponding audio message [2].

1.2 Speaker Verification

Speech is the most convenient communication tool for humans. Speech still remains as an important means of communication. Speaker verification is a biometric modality that uses an individual's voice for recognition or verification purpose. It is a well-known fact that the humans outperformed machine in the case of speaker recognition as the humans are more robust than machine to distortion and noise and is proved in [9] that the machines has the grater ability than humans to distinguish between the voices of

identical twins. Some of the potential applications of Speaker verification includes, i) Telephone based verification systems have a number of applications, particularly in transactions requiring secure access to financial information, teleshopping, telebanking etc. ii) Fast and secure physical access to restricted locations, iii) electronic commerce, iv) to trace the voice of a person in VoIP. The speaker verification system plays a major role in tracking the speech or conversation of a wanted person for security / crime related cases in various media such as telephones, mobile phones, VoIP, radio and TV and so on in real time mode or offline mode by analyzing the stored data.

Speaker verification has been an interesting area of research for many years. Many researches have been done and some of them have reached high performance level. Many techniques have been proposed for speaker verification systems including Dynamic Time Wrapping (DTW), Hidden Markov models (HMM), Artificial Neural Networks (ANN) and Vector Quantization (VQ).

Some of the merits of speaker verification are as follows:

i) The cost of implementation is low because there is no special hardware required.
ii) Speaker Verification is easy to use and it has a very high user acceptance rate.
iii) Voice biometrics is the only biometric that allows users to authenticate remotely.
iv) It is quick to enroll in a Speaker Verification system.

The task of speaker verification is a subset of the general problem of speaker recognition, which includes the task of speaker identification. Speaker identification means comparing an unknown voice as one of a set of known voices, whereas speaker verification means determining whether an unknown voice matches the known voice of a speaker whose identity is being claimed. The speaker verification task involves a binary comparison, its performance is independent of population size but the speaker identification task involves $\boldsymbol{N}$ + 1 decision for a population size of $\boldsymbol{N}$ speakers which degrades the performance when the size of the population increases.

Speaker-specific characteristics of speech are due to differences in physiological and behavioral aspects of the speech production system in humans. Hence, there are large variability in the speech signal between speakers and, more importantly, between speech data collected from the same speaker at different times which is termed as intra-speaker variability, one of the major issue in speaker verification. The features used in any speaker verification system should be discriminated between speakers while being tolerant of intra-speaker variability, should be easily measurable from the speech signal, should be stable over time, and not be susceptible to mimicry by impostors.

The robust speaker recognition system faces a variety of challenges for identifying or verifying speaker identities in noisy environments, which cause a large range of variance in feature space and therefore are extremely difficult for statistical modeling. To compensate this effect, and also considering the capacities and limitations, normalization techniques at the score level and transformation techniques at the feature level are discussed in [8] and some of the recent advances in speaker verification are also introduced.

Though speaker recognition does not offer the same robustness and precision than other biometric traits such as fingerprint and iris, strong efforts are being done to enhance the performance, due to its particular set of characteristics that can permit to manage some vulnerability attacks [7]. The crossover ratio of various biometric technologies is given in Table 1. It is found from the table that voice biometrics has a higher cross over ratio than the other technologies available. It is not appropriate to use them independently for authorization to systems that require high security.

Table 1 Crossover Rates for Various Biometrics [28]

Biometric accuracy	**Crossover**
Retinal scan (%)	0.0000001
Iris Scan (%)	0.000763
Fingerprints (%)	0.2
Hand Geometry (%)	0.2
Signature Dynamics (%)	2
Voice dynamics (%)	2

The effectiveness of various speaker specific features has been studied in [22]. Voice pitch has been used in various speaker verification applications but it is not always easy to measure, especially in noisy environments and it is easy to mimic [1]. Voice pitch patterns change significantly due to stress and speech effort levels. Formant frequencies contain speaker specific information and can be used to distinguish between users, but the main drawback is the difficulty in measuring the formant frequencies. Though several methods of extracting features exists, the choice of a particular representation is determined by practical considerations, such as, ease of computations, storage requirements, methods of pattern matching, susceptibility to channel distortions etc.
A new approach to text-independent phoneme based speaker verification is used in [23], in which a two-stage classifier is used. The first stage consists of a speaker-independent phoneme detector trained to recognize a phoneme that is distinctive from speaker to speaker. The second stage is trained to recognize the frames of speech from the target speaker that are admitted by the phoneme detector. The study indicated that the phoneme-based approach helps in eliminating the rejection of target speaker, but to identify the phoneme specific to a particular speaker is a difficult process. A speaker verification system described in [3] is based on a

vector quantization (VQ) approach that incorporates dynamic time warping (DTW), cohort models, and a discriminator to separate the true speakers and imposters, which achieves an equal error rate (EER) of 0.92% when the true speaker and the imposters spoke different pass phrases and 4.30% when they spoke the same phrase.

The influence of cepstrum parameters on text independent speaker verification and speech recognition has been investigated in [19] and the influence of formant frequencies on the efficiency of the speaker verification system shows a better performance than the mel scale. A MAP adaptation algorithm for the VQ model is derived in [10] as a special case of the MAP adaptation for GMM, involving only the centroid vectors in which the VQ approach achieves speed-up in training compared to GMM with comparable accuracy.

To achieve both computational efficiency and high accuracy in text-independent speaker verification, an integrated system with structural Gaussian mixture models (SGMMs) and a neural network is proposed in [4] and a 5% relative reduction in equal error rate (EER) is achieved by this method. Though the computational cost is reduced significantly, it is achieved by scoring only a subset of the Gaussian components during verification.

Even though vector quantization speaker modeling was popular in 1980s and 1990s [11], [24], GMMs became the dominant approach in text-independent speaker verification [20]. However, in a GMM-based text-independent speaker verification system, generally a universal background model (UBM) with a large number of Gaussian mixture components is created based on hours of speech data from non target speakers which results in more processing time during verification. A frame-by-frame comparison between speaker model and background model based on a normalization framework for HMM-based text-dependent speaker verification gives an alternative approach, which leads to similar performances with respect to classical HMM-based approach [5]. It shows a significant improvement in error rates though the combination of acoustical and alignment scores is marginally less. Though the results obtained in [12] shows an improved performance compared to other existing techniques, the number of iteration is more to identify the speaker specific information from the high frequency components, which may increase the time complexity, and also this may not be considered in the case of large population.

Artificially mixed target and masker speech utterances are employed to improve the performance of speaker recognition system [14]. It has been observed from the results that, when the target and masker utterance have the same gender, the recognition system has a performance at 0 dB equal to that of humans; in other conditions the error rate is roughly twice the human error rate. But, in this case, the adaptation and training of clean speech models from noisy speech data remain challenging problem.

Speaker authentication systems based on the combination of several speaker classifiers [15] showed that the utterance verifier performs worse than the speaker verifier, but the combination of the two verifiers shows an improved performance when a good compromise is made between performance, complexity and adequate use of the training material. To compensate for the channel distortion, a robust speaker recognition method based on position-dependent Cepstral Mean Normalization (CMN) depending on the speaker position achieved a relative error reduction rate of 64.0% and 30.2% from the position-independent CMN [16]. Speaker clustering methods for speech recognition based on vocal tract (VT) size results in higher recognition rates than conventional speaker clustering methods based on acoustic criteria, but an increased amount of training data is used to improve the robustness [17]. Threshold setting is one of the major issues in speaker verification on which a preliminary investigation in setting a prior threshold for speaker verification has been done in [26].

The accuracy of speaker verification is affected by several factors. The difference in telephones can be problematic. Background noise, illness, and vocal changes from age can all affect accuracy. Accuracy of speaker is measured in three categories:

(i). Failure to enroll-user's registration into the system is not successful.
(ii). False acceptance-user is authenticated when he/she should not be.
(iii). False rejection-user is not authenticated when he/she should be.

To handle the general and specific techniques from variable sources in automatic speaker recognition, speech analysis and acoustic modeling is used [1]. The entropy measure based speaker verification gives a better performance by setting prior threshold but only a preliminary stage of investigation in information based speaker verification [25]. To perform text independent speaker verification by constructing the background speaker models and by using speaker-clustering method a preliminary research work has been done which gives a better result [13]. Computation cost may be reduced if a combination of maximum likelihood posteriori (MLP) and Gaussian Mixture Model (GMM) are used in text independent speaker verification [3]. A fuzzy c-means clustering method gives a very good result in radio environments [21].

The state-of-art speaker verification models are based on Hidden Markov Model (HMM), Gaussian Mixture Model (GMM, Vector Quantization (VQ), Dynamic time Warping (DTW), Neural Networks, Support Vector Machine (SVM) etc. The most difficult problem in speaker verification is the intra speaker variability which is examined in this paper by setting training the individual user and by computing the threshold of the speaker to achieve a better performance. Briefly, the task involved in this systems is to collect sample from speakers, after the preprocessing and feature extraction steps, codebooks are generated using k-means algorithm to generate speaker models and to compute threshold for each user. Here a closed-set of 90 speakers are used. Training is given to individual users incrementally. Whenever a target speaker is to be verified, in the previous cases, the speaker who is very close to the target speaker is selected. In this

current work, best two matches of the target speaker are listed out and then the overall scoring recognizes the potential target speaker. This paper reviews the performance of this approach and examines the similarities and difference between the results obtained using Vector Quantization method, Gaussian Mixture model methods and the Hidden Markov Model and those obtained by listeners. The current paper examines the reasons for speaker verification failures and as a result of this analysis proposes a novel technique to produce a new and improved result on the speaker verification challenges.

1.3 Speech Recognition

In recent years, Automatic Speech Recognition (ASR) has reached very high levels of performance, with word-error rates dropping by a factor of five in the past five years. This current state of performance is largely due to improvements in the algorithms and techniques that are used in this field. As a result, the accuracy level of ASR systems is improved especially when using a combination of various algorithms and techniques. The major issue in ASR system is noise interference - Signal to Noise Ratio. Feature extraction is the initial step in ASR. The most frequently used parameters for feature Extraction are pitch, formant frequency and bandwidth, Linear Predictive Coefficients (LPC), Linear Predictive Cepstrum Coefficients (LPCC), Mel-Frequency Cepstrum Coefficients (MFCC) and so on. The formant, LPC and LPCC are related to vocal tract, and are good user identification characteristics with high SNR (signal to noise ratio). However, when the SNR is low, the differences between the vocal tract parameters estimated from noisy speech signal and those of the real vocal tract model are big. Thus, these characteristic parameters cannot correctly reflect user's vocal tract features. The MFCC parameter mainly describes the speech signal's energy distribution in a frequency field. This method, which is based on the Mel frequency scale and accords with human hearing characteristics, *has better anti-noise ability* than other vocal tract parameters, such as LPC [27]. The next step is pattern matching, in which, the word recognition requires the comparison between the entry signal of the word and the various words of the dictionary. The problem can be solved efficiently by a dynamic comparison algorithm whose goal is to put in optimal correspondence the temporal scales of the two words. The technique of Dynamic Programming is a powerful method for isolated spoken-word recognition.

The problem of speech recognition belongs to a much broader scientific topic called pattern recognition or pattern matching/classification. But this faces the computational complexity and time consumption issue. Vector Quantization is comparatively easier to implement and matching time is less [6]. VQ is a general class of methods that encode groups of data rather than individual samples of data in order to exploit the relation among elements in the group to represent the group as a whole more efficiently than each element by itself. The similarity or distortion measure is an advantage of VQ algorithm since it has a built-in distance measure in its computation process.

2. METHODOLOGY

The Voice Activated E learning system is designed in such a way that would guide and instruct visually impaired users to accomplish their learning task successfully. In general, this voice activated system was developed as two main subsystems namely the user verification subsystem and isolated word recognition subsystem. The user verification subsystem was used to authorize the user, and the word recognition subsystem was developed for the purpose of command identification. These two subsystems were combined to obtain the E learning system.

The methodology carried out is described below

1. Collection of speech samples and voice models from different users.
2. Extracting distinguished and discriminative word specific and voice specific features from the collected samples and producing a set of feature vectors.
3. Training the feature vectors in order to build unique voice models.
4. Matching/testing unknown feature vectors against the trained voice models in order to obtain the accuracy/recognition rate.
5. Verifying the user and recognizing the word uttered by the user, thereby responding appropriately.
6. Evaluating the E-learning system.

2.1 User Enrollment

Firstly, the system must record the user's name and record the user's voice. Then it generates a User Specific Threshold (UST). Secondly, the system must record a specific set of command words in the user's voice and allow users to enroll for the subjects of interest, from a provided set of subjects by providing a specific set of subject keywords in the user's voice and prompts the user to complete the entire enrollment process.

2.2 User Login

In the user login session the system must instruct user before starting the Login process. It captures the user's name and voice sample and verifies the user's claim and provides the verdict. It also provides a set of chances to an unauthorized user, before exiting the application. It allows the user to enter into the next phase, once his/her claim is accepted.

2.3 Pace Selection

Here the system provides options to choose the speed at which the audio is played out (*ie slow, medium, fast.)* and check for the set of subjects enrolled by the user, to move to the corresponding page.

The system captures the subject name uttered by the user and presents the set of subjects enrolled by the user, if it is more than one through audio as well as in text form and list down the set of topics available for learning in a particular subject,

the topics the user has already covered and captures the topic name uttered by the user.

Then a short introduction on the lesson for the session is given and waits for the user to instruct, before starting the lessons. It also provides the user the options to move forward and backward through the chapters and allow the user to listen to a particular content any number of times. It also provides alternate chapters for better understanding of the concepts and plays the audio at the requested rate. It informs the user on finishing all the chapters in a topic and requests the user to choose another topic, on completion of a topic.

2.4 Challenges

The visually impaired person cannot learn in way the normal person can learn. They may not be convenient with software and hardware used by the normal person and may have difficulty in receiving and interpreting output from the computer; giving commands or entering data into the computer; also they should have sufficient computer knowledge.

2.5 Transforming GUIs into SUIs

Since one of the goals of this work is to enable speech access to the E learning environment, our initial SUI (Speech User Interface) designs were influenced by the existing graphical interfaces. The evolution of our SUI design shows a clear trend towards interpersonal conversational style and away from graphical techniques

2.5.1 Content Development

The content is presented in such a way that many examples have been included for the ease of understanding. The audio contents are being played in a brief, and a clear manner.

2.5.2 Ambiguous Silence

Another speech-related problem is the difficulty users have in interpreting silence. Sometimes silence means that the speech recognizer is working on what they said, but other times, it means that the recognizer simply did not hear the user's input. A solution to this problem can be obtained by utilizing keyboard controls, to inform the system that the user has uttered his command.

2.5.3 Pacing

A very important aspect of content learning for the visually impaired users is the pace in which the audio is presented to the user. The user is provided with a set of pace options. Variable Audio output rate allows content to be played to suite the individual's learning ability.

2.6 Architecture diagram

Figure 1 shows the architecture of the voice activated elearning system for the visually impaired.

2.7 Implementation

2.7.1 User verification subsystem

The most prevalent and dominant method used to extract features is calculating Mel-Frequency Cepstral Coefficients (MFCC). This is done with the help of framing, in which the speech samples are blocked into frames of N samples, with adjacent frames being separated by M (M<N). Signal discontinuities at the beginning and end of each frame are minimized by tapering the signal to zero at the ends. If the window is defined as

w (n), 0<=n<=N-1,

Where N is the number of samples in each frame, then the result of the windowing is the signal as in Equation 1,

$$y(n)=x(n)w(n) \quad \text{----- (1)}$$

$$w(n)=0.54-0.46\cos(2\prod n/N-1), \quad 0<=n<=N-1 \quad \text{----- (2)}$$

The result of equation 2 is called Spectrum. The next step is calculating the fast Fourier transform (FFT). Each frame of N samples is converted from the time domain into frequency domain. FFT is a fast algorithm to implement Discrete Fourier Transform (DFT) which is defined on the set of N samples $\{x_n\}$ as in Equation 3

$$Xn = \sum x_k \, e\text{-}2j\prod n/N, \; n=0,1,2\ldots\ldots N-1 \quad \text{----- (3)}$$

The mapping between frequency in Hertz and the Mel scale is linear below 1000Hz and logarithmic above 1000Hz.The Mel frequency can be computed from the raw acoustic frequency using Equation (4)

$$Mel(f)=2595*\log10(1+f/700) \quad \text{------ (4)}$$

In the final step, log Mel spectrum is converted back to time using Discrete Cosine Transformation (DCT) as in Equation (5) and the result is called the Mel frequency cepstrum coefficients (MFCC).

$$Cn = \sum(\log S_k)\cos[\, n(k-1/2)\prod/K], \quad \text{----- (5)}$$

Where, n=1,2.....K , k=1,2.....K

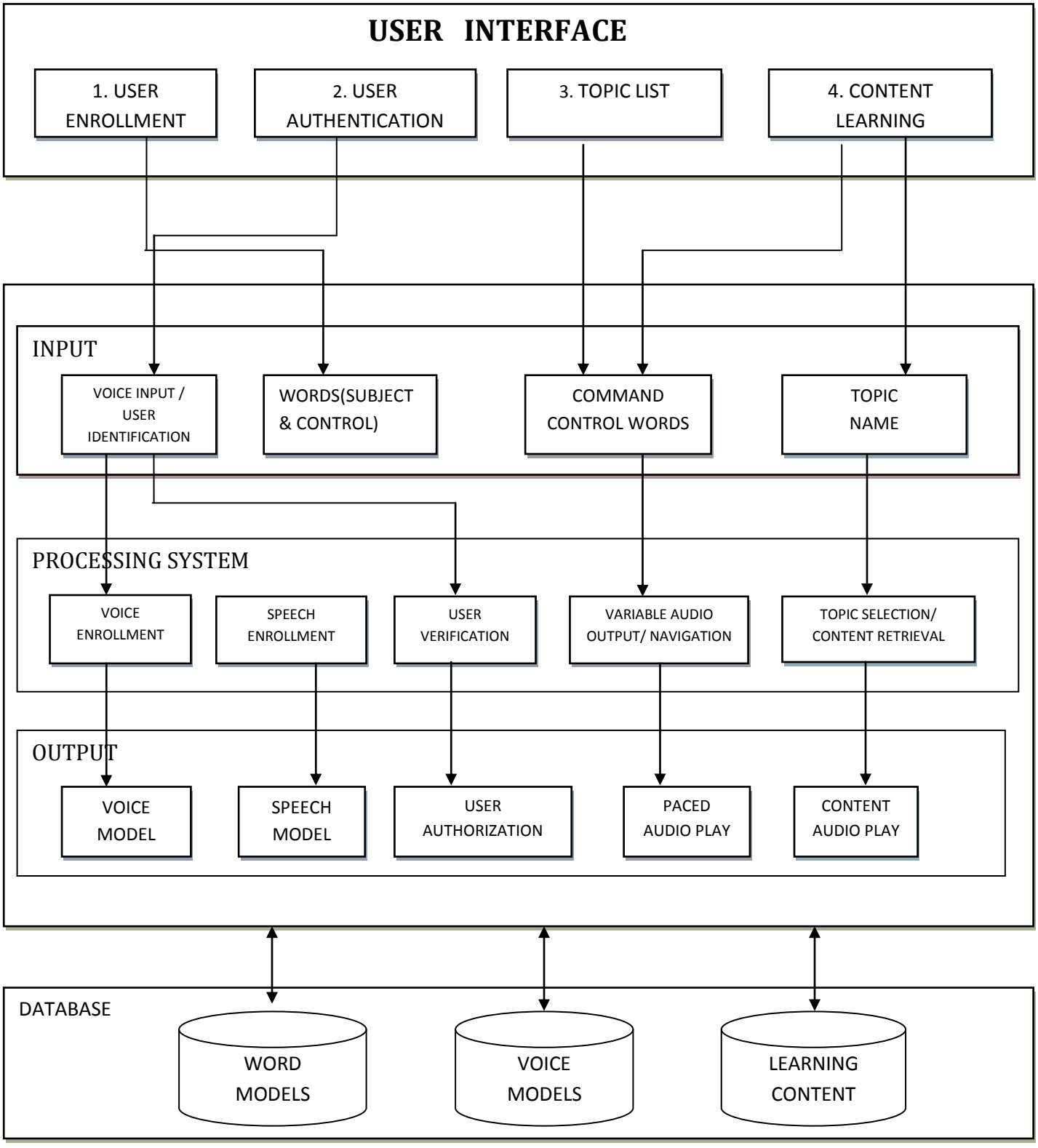

Figure 1 Architecture for Voice activated e-learning system

Feature training is done using vector quantization (VQ) which is the process of taking a large set of feature vectors and producing a smaller set of feature vectors that represent the centroids of the distribution, i.e. points spaced so as to minimize the average distance to every other point and k-means clustering algorithm.

Here the input is
$I=\{i_1,i_2……..i_n\}$
the set of data items to be clustered, and k is the number of clusters.
The output obtained is $C=\{c_1,c_2…….c_k\}$
the cluster centroids and
m = I->C, the cluster membership.
The matching of an unknown voice is performed by measuring the Euclidean distance between the features vector of the unknown voice pattern to that of the voice models (voice model) of the available voice patterns in the database.

Identity of unknown word is given as

$$d(x,y) = \sqrt{\sum_{i=1}^{D} (x_i - y_i)^{\wedge}2} \quad \text{---- (6)}$$

Where x_i is the i^{th} input features vector, y_i is the i^{th} features vector in the voice model, and d is the distance between x_i and y_i.

2.7.2 Speech recognition subsystem

The most prevalent and dominant method used to extract features is calculating Mel-Frequency Cepstral Coefficients (MFCC). Dynamic programming technique is used here. Given the set of word templates and the uttered word, the word that is closest to the uttered word will be found by comparing the spoken word with all the templates and choosing the one that has the minimum distance (similar to) with the uttered word.

2.7.2.1 Voice Enrollment

The user enters his name and records his Voice Sample. Features are extracted and a voice model is created from the given voice sample.

2.7.2.2 Speech Enrollment

The user is prompted with a keyword and records the keyword, with speech input. Features are extracted and a voice model is created from the given voice sample.

2.7.3 Command recognition subsystem

The user utters the name of the subject and the system extracts the features from the voice sample. The word pattern obtained is compared with the patterns stored in the database and the appropriate word is retrieved. The word obtained, is played back to the user, and then passed on to the Content Retrieval subsystem.

2.7.4 Content Retrieval subsystem

Here the list of available topics is given for selection and the System obtains the relevant audio material and plays it to the user. The user is provided with two or more speed options to choose the audio output rate.

3. EXPERIMENTAL RESULTS

The experimental results show the relevant testing results and discussions on the findings of this work. This shows the testing performed and the experimental results starting from the User verification subsystem, speech recognition subsystem followed by the overall performance of the E learning system.

3.1 Performance of voice verification subsystem

A Verification system was designed to verify a user, given his voice input and his identity claim. The system verifies the input sample by:

1) Matching with the given user's Identity.
2) NOT matching with the given user's Identity.

In the User Verification subsystem 2 tests are evaluated: Test 1 for evaluating the true input and Test 2 for evaluating the false input

Test 1: Evaluating True Input

POSITIVE TESTING: Given a positive user Person 'A' , to check if the system identifies the user as person 'A'.

User: Person 'A' is an enrolled User.
Testing: To check if the system authorizes the user.

Total no. of Users:	50
No. of users identified as per claim:	48
True Positive:	96%
False Negative:	4%

Figure 2 shows the result of the true positive and false negative of a true input in a speech recognition subsystem.

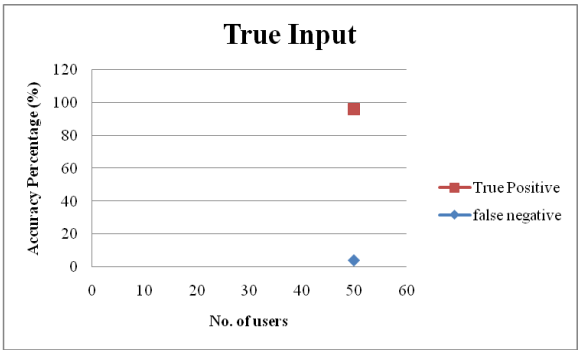

Figure 2 Speech Recognition Subsystem (True input)

Test 2: Evaluating False Input

NEGATIVE TESTING: Given a negative user, Person 'A', to check if the system identifies the user Person 'A'.

USER: Person 'A' is an unknown User (not enrolled).

TESTING: To check if the system authorizes the user.
Total no. of Users: 50.
No. of users authorized as per claim: 3
True Negative: 94%.
False Positive: 6%.

Figure 3 shows the result of the true negative and false positive of a true input in a speech recognition subsystem.

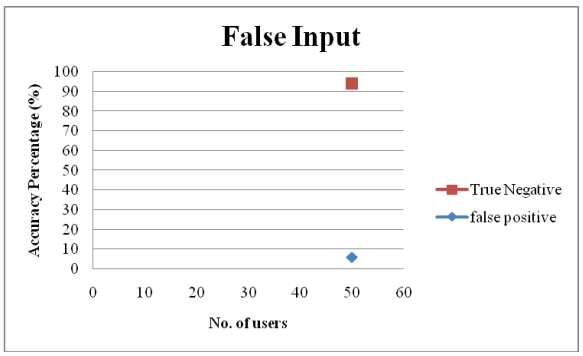

Figure 3 Speech Recognition Subsystem (False input)

3.1.1 Results of user verification

Accuracy: 96%
Test Input: 6 seconds
Verification Time: < 1 second
Disk Usage space: 94.4 KB (on an average for each user data)

3.2 Performance of Speech Recognition Subsystem

The focus of this evaluation was recognizing an isolated distinct word from a known set of distinct words. A set of ten distinct words was considered as speech units. Table 2 gives the sample set of words and the results of the speech recognition subsystem. The system was tested with 50 users belonging to different ages and genders. Each distinct word was recorded by all the users and their recordings were saved for further processing. The number of speech samples per word was just one. The system was then tested with the users again uttering one amongst the nine distinct words and if the system identified the correct word or not. Each user is instructed to record ten distinct words. To check if the system identifies the word the user utters. The identification accuracy of the words of the system for 45 users is found to be 90%.

Table 2 Sample of Speech Recognition subsystem

Serial number	Word	correct matching (no of users)	*Wrong matching (no of users)
1.	Slow	45	5
2.	Normal	39	11
3.	Fast	46	4
4.	Subject	40	10
5.	Project	38	12
6.	English	49	1
7.	Mun	50	0
8.	Aduthadu	50	0
9.	Noun	48	2
10.	Phrase	40	10
Total	**-**	**445**	**55**

POSITIVE IDENTIFICATION:
Total Number of Users: 50
90% Accuracy Achieved With: 45
Performance of at least 90% accuracy: 90%
Performance accuracy: 80%

The above result was obtained for 8 users.

POSITIVE IDENTIFICATION:
Total Number of Users: 50
90% Accuracy Achieved With: 45
80% Accuracy Achieved With: 5
Performance of at least 80percent: 100%

Figure 4 shows the results of the performance of word recognition in a speech recognition subsystem.

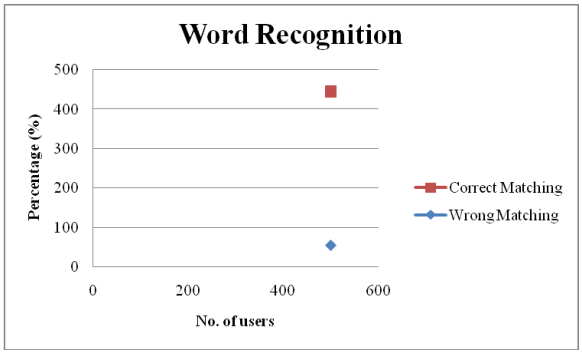

Figure 4 Performance of word recognition

3.2.1 Results of Speech Recognition Subsystem

- There were 50 users, uttering 10 words each, and the results were presented for a total of 500 utterances.
- 45 users had at least nine words identified correct. Thus 405 correct utterances.
- 5 users had eight words identified correct. Thus 40 correct utterances.

Total number of Utterances:	500
Total Correct Matching:	445
Total Wrong Matching:	55
Accuracy:	89%

4. PERFORMANCE OF E LEARNING SYSTEM

The E learning system consists of a user enrollment, user authentication and lots of word recognition processing. For the purpose of testing the system, 20 users were enrolled and were asked to use the system.

4.1 User Enrollment

The following details are provided by the user to the system:

1. User Voice

Recording Time: 6(voice) + 6(confirm) = 12 seconds
Disk space: 96 KB

2. Command Words

Recording Time: 60 seconds
Disk space: 400 KB (approx)
Overall Enrollment Time: 300 seconds (approx)
Overall Disk storage: 500 KB (approx)

4.2 User Authentication

For the purpose of authorizing a user, we have made use of the User Verification subsystem.

The user utters his name and gives his voice sample for six seconds. The system verifies the user. The user is provided 'Three Chances' to enter into the system, before the application exits.

Performance: 95% accuracy

4.3 Command Recognition

The system consists of a set of command words belonging to both the English and the Tamil vocabulary.

1. Pace Command (6): Used for choosing the required pace for learning.

- Slow, Normal, Fast, Mella, Midhamana, and Vegamaga.

2. Navigation Command (8): used to navigate pages.

- Start, Next, Previous, Replay, Thodangu, Aduthadu, Mun sel, Marubadiyum

3. English subject Words (12) : to select the subjects

- English, Adjective, Adverb, Basics, Conjunction, Modals, Noun, Participle, Phrase, Preposition, Pronoun, Tense

4. Tamil subject words (6): to select the subjects

- Tamizh, Adakam udaimai, Kadavul vazthu, Ozhukam udaimai, Pugazh udaimai, Sei nandri aridhal.

Complete Performance of the E-learning system:

95% (authentication) + 92.1% (word recognition)
= 93.55 % accurate.

Figure 5 shows the results of the performance of user authentication and command recognition in a speech recognition subsystem.

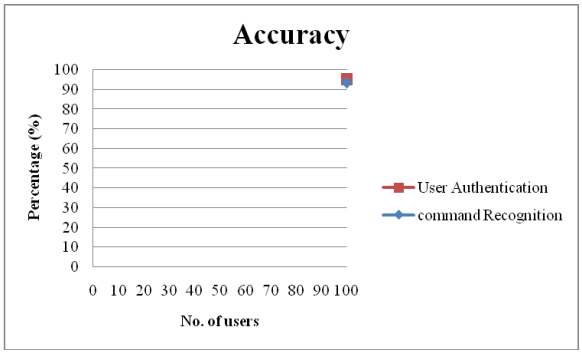

Figure 5 Performance of user authentication and command recognition

5. CONCLUSION AND FUTURE WORK

This project has achieved and met the objectives of developing it, and it is hoped that this project will benefit the end users as it is designated for a purpose. The voice activated E-learning system could enroll users, verify their voiceprint, and provided a speech interface for users to learn. This system allows users with language preferences though the system is mainly in English language. Performance of this system is good with an accuracy of 95% in User Authentication subsystem and 92.1% in the Speech recognition subsystem and the overall accuracy of 93.55% in the entire system. This system could be implemented in many academic institutions, for educating visually impaired children. Overall, the project was a success with the basic requirements being satisfied.

- The system shall provide more language preferences for users, so that the most commonly used and spoken languages could be provided in the system, which will ease the process of communication.
- The system shall be integrated with "Emotion based Interactive and adaptive E learning system" where the contents change based on the facial reactions given by the user. The authentication can be made much stronger by making use of user's facial features and voice parameters.

6. ACKNOWLEDGEMENT

This work is supported by the NTRO, Government of India. NTRO provides the fund for collaborative project "Smart and Secure Environment" and this paper is modeled for this project. Authors would like to thank the project coordinators and the NTRO members.

7. REFERENCES

[1] Atal .B. S (1972), Automatic Speaker Recognition Based on Pitch Contours, International Acoustics Society, America, vol. 52, pp. 1-687.

[2] Athanasios Drigas, Leyteris Koukianakis, Yannis Papagerasimou, "An E-Learning Environment for Nontraditional Students with Sight Disabilities", 36th

ASEE/IEEE Frontiers in Education Conference, IEEE 2006.

[3] Belfield .W. R. and Mikkilineni .R. P. (1997), Speaker Verification Based On A Vector Quantization Approach That Incorporates Speaker Cohort Models And A Linear Discriminator, IEEE, Vol.1, 4525-4529.

[4] Bing Xiang, Toby Berger (2003), Efficient Text-Independent Speaker Verification with Structural Gaussian Mixture Models and Neural Network, IEEE Transactions on Speech and Audio Processing, Vol. 11, No. 5, IEEE.

[5] Charlet.D, Jouvet.D, Collin.O (2000), An Alternative Normalization Scheme in HMM-Based Text-Dependent Speaker Verification, Speech Communication, Elsevier publications, Vol. 31, pp. 113-120.

[6] Clarence Goh Kok Leon, Politeknik Seberang Perai, Jalan Permatang Pauh, "Robust Computer Voice Recognition Using Improved MFCC Algorithm", International Conference on New Trends in Information and Service Science, IEEE 2009.

[7] Enric Monte-Moreno, Mohamed Chetouani, Marcos Faundez-Zanuy, Jordi Sole-Casals (2009), Maximum Likelihood Linear Programming Data Fusion for Speaker Recognition, Speech Communication, Vol. 51, pp. 820-830.

[8] France and Janez, (2008), A textbook on Speech Recognition Technologies and Applications.

[9] George R. Doddington, Mark A. Przybocki, Alvin F. Martin, Douglas A. Reynolds (2000), The NIST Speaker Recognition Evaluation-Overview, Methodology, Systems, Results, Perspective, Speech Communication, Vol. 31, pp. 225-254.

[10] Gurmeet Singh, Ashish Panda, Sauruv Bhattuchulyya, Thambipillui Stikunthun (2003), Vector Quantization Techniques for GMM Based Speaker Verification, ICASSP 2003, IEEE 2003, pp. 65 - 68.

[11] He, J., Liu, L., Palm, G. (1999), A Discriminative Training Algorithm for VQ-Based Speaker Identification, IEEE Transactioons, Speech Audio Process, 7 (3), 353-356.

[12] Hemant Misra, Shajith Ikbal, B. Yegnanarayana (2003), Speaker-Specific Mapping for Text-Independent Speaker Recognition, Speech Communication, Vol.39, pp 301-310.

[13] Hou Fenglei, Wang Bingxi (2002), Text Independent Speaker Verification Using Speaker Clustering and Support Vector Machines, the Proceedings of ICSP'02, IEEE 2002, pp 456-459.

[14] Jon Barker a, Ning Maa, Andre´ Coy a, Martin Cooke (2010), Speech Fragment Decoding Techniques for Simultaneous Speaker Identification and Speech Recognition, Computer Speech and Language, Vol.24, pp 94-111.

[15] Leandro Rodrguez-Linaresa, Carmen Garca-Mateob, Jose Luis Alba-Castrob (2003), On Combining Classifiers for Speaker Authentication, Pattern Recognition, Vol. 36, pp 347-359.

[16] Longbiao Wang, Norihide Kitaoka, Seiichi Nakagawa (2007), Robust Distant Speaker Recognition Based on Position-Dependent CMN by Combining Speaker-Specific GMM With Speaker-Adapted HMM, Speech Communication, Vol.49, pp. 501-513.

[17] Masaki Naito, Li Deng, Yoshinori Sagisaka (2002), Speaker Clustering for Speech Recognition Using Vocal Tract Parameters, Speech Communication, Vol. 36, pp. 305-315

[18] Mohammad Nouman, Nasir Naveed, Muhammad Abdul Basit Khan,"E-Learning for Visually Impaired", University of Pakistan, 2006.

[19] Paul Cristea and Zica Vrilsan (1999), New Cepstrum Frequency Scale For Neural Network Speaker Verification, IEEE, vol. 1, pp. 573-576.

[20] Reynolds, D., Quatieri, T., Dunn, R.(2000), Speaker Verification Using Adapted Gaussian Mixture Models. Digit. Signal Process, 10 (1), pp.19-41.

[21] Robert Batusek and Ivan Kopecek, "User Interfaces for the Visually Impaired people", Masaryk University, 2000.

[22] Rosenberg A.E, "Automatic Speaker Verification: A Review," *Proc. IEEE*, vol. 64, pp. 475-487, Apr. 1976.

[23] M. Savic J. SoFensen 91992), Phoneme Based Speaker Verification, IEEE, Vol. 2 pp. 165-168.

[24] Shung-Yung Lung (2007), Efficient Text Independent Speaker Recognition with Wavelet Feature Selection Based Multilayered Neural Network Using Supervised Learning Algorithm, Pattern Recognition, Vol. 40 pp. 3616-3620.

[25] Soong, F.K., Rosenberg, A.E., Juang, B.H., Rabiner, L.R. (1987), A Vector Quantization Approach to Speaker Recognition. AT&T Tech. J. 66, pp.14-26.

[26] Tuan Pham, Michael Wagnel (2000), Information Based Speaker Verification, IEEE, pp. 278-281.

[27] Wang Yutai, Li Bo, Jiang Xiaoqing, Liu Feng, Wang Lihao, "Speaker Recognition Based on Dynamic MFCC Parameters", IEEE 2009.

[28] Yi-Hsiang Chao, Wei-HoTsaic, Hsin-MinWang, Ruei-ChuanChang (2009), Improving the Characterization of the Alternative Hypothesis Via Minimum Verification Error Training with Applications to Speaker Verification, Pattern Recognition, Vol. 42, pp. 1351 - 1360.