Research Journal of Applied Sciences 2 (6): 677-684, 2007

Development of a Text Dependent Speaker-Identification Security System

A.O. Afolabi, A. Williams and Ogunkanmi Dotun
Department of Computer Science and Engineering,
Ladoke Akintola University of Technology, Ogbomoso, Nigeria

Abstract: Different methods of providing security of sensitive/personal data have been developed which include Personal Identification Numbers (PINs), Passwords, Identity Cards, etc. and these have not been successful enough in providing the required security. Passwords can be easily forgotten or stolen, ID cards can be forged and this stresses the need for improving security by the use of the voice, which is unique to each person and cannot be stolen and is very difficult to forge. The voice among other biometric features is the safest level of protection of sensitive/personal data, therefore, this research provide a framework of a text dependent speaker identification system which serves as a security tools in system operations so that by identifying the voice of a user there cannot be any denial of transactions within the system.

Key words: Text dependent, speaker identification, security system development, PINs, ID

INTRODUCTION

Biometrics is automated methods of recognizing a person based on a physiological or behavioral characteristic. Among the features measured are face fingerprints, hand geometry, handwriting, iris, retinal, vein and voice. Biometric technologies are becoming the foundation of an extensive array of highly secure identification and personal verification solutions (Moshe, 2002). As the level of security breaches and transaction fraud increases, the need for highly secure identification and personal verification technologies is becoming apparent.

Biometric-based solutions are able to provide for confidential financial transactions and personal data privacy. The need for biometrics can be found in federal, state and local governments, in the military and in commercial applications. Enterprise-wide network security infrastructures, government IDs, secure electronic banking, investing and other financial transactions, retail sales, law enforcement and health and social services are already benefiting from these technologies in major developed countries of the world.

Speaker recognition is the process of automatically recognizing who is speaking based on individual information included in speech waves. This technique makes it possible to use the speaker's voice to verify their identity and control access to services such as voice dialing, banking by telephone, telephone shopping, database access services, information services, voice mail, security control for confidential information areas and remote access to computers (Becchetti and Ricotti, 1999).

TEXT-DEPENDENT VS. TEXT-INDEPENDENT SPEAKER RECOGNITION

Speaker recognition systems are classified as text-dependent (fixed-text) and text-independent (free-text). The text-dependent systems require a user to re-pronounce some specified utterances, usually containing the same text as the training data. There is no such constraint in text independent systems. In the text-dependent system, the knowledge of knowing words or word sequence can be exploited to improve the performance (Markowitz, 2002).

There are a few methods that are used for speaker verification. The text dependent speaker recognition methods can be classified into DTW (Dynamic Time Warping) or HMM (Hidden Markov Model) based methods.

Text-independent speaker verification has been an active area of research for a long time because performance degradation due to mismatched conditions has been a significant barrier for deployment of speaker recognition technologies.

Figure 1 shows the basic structures of speaker identification and verification systems.

Corresponding Author: A.O. Afolabi, Department of Computer Science and Engineering,
Ladoke Akintola University of Technology, Ogbomoso, Nigeria

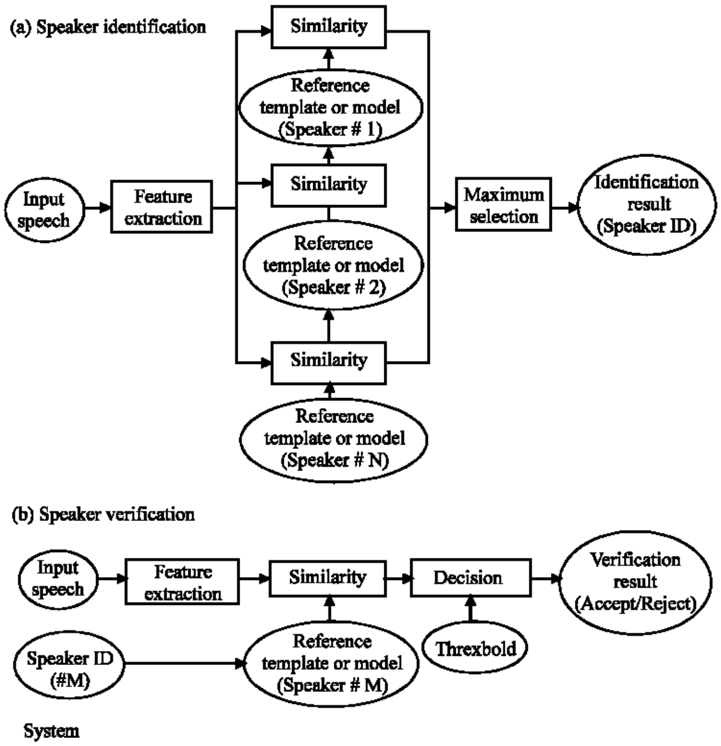

(a) Speaker identification
Input speech
Feature extraction
Similarity
Reference template or model (Speaker # 1)
Similarity
Reference template or model (Speaker # 2)
Similarity
Reference template or model (Speaker # N)
Maximum selection
Identification result (Speaker ID)
(b) Speaker verification
Input speech
Feature extraction
Similarity
Decision
Verification result (Accept/Reject)
Speaker ID (#M)
Reference template or model (Speaker # M)
Threxbold
System

Fig. 1: Basic structures of speaker recognition systems

HOW VOICE VERIFICATION WORKS

Voice biometrics works by digitizing a profile of a person's speech to produce a stored model voiceprint, rather like a template, which is referred to each time that person attempts to access secure data (Jason, 2003). The position and movement of the glottal tissues, lips, jaw and tongue correspond with speech movements in the vocal tract. Biometrics technology reduces each spoken word into segments: Sub-word like syllables, phonemes, triphones or similar units of sound, composed of several dominant frequencies called formants, which remain relatively constant over that segment. Each segment has 3 or 4 dominant tones that can be captured in digital form and plotted on a table or spectrum. This table of tones yields the speaker's unique voice print.

The voice print is stored as a table of numbers, where the presence of each dominant frequency in each segment is expressed as a binary entry. Since all table entries are either 1's or 0's, each column can be read bottom to top as a long binary code. When a user attempts to gain access to protected data, their pass phrase is compared to the previously stored voice model and all other voiceprints stored in the database. Each speech sound in the user's pass phrase is queried in an anti-speaker database. Since some characteristics of a person's voice are the same as another's, the system authenticates the user by comparing the user's common features with those in the anti-speaker database and eliminating those common elements from the sample to be authenticated. When all features matching others are removed, the system is left with only the unique features of the user's voice. These unique features, compared with the enrolled pass phrase, are the characteristics, which determine successful authentication.

Voice verification scores are controlled by setting a threshold of reliability or acceptance. Different businesses will have different priorities in terms of the acceptable confidence or matching level. Each organization's administrator is therefore able to set a specific acceptable statistical score. If the threshold score is higher than the preset level, the user is accepted. If it is lower, the speaker is denied access.

When authenticating, a user is asked to answer up to 3 prompted questions, the answers to which are easily remembered by the user. In order to provide audible content of at least one second in length, typical prompts are:

- User's first, middle and last name
- User's date and month of birth
- Mother's first, middle and last maiden name
- Home telephone number

SPEAKER VERIFICATION ALGORITHMS

The most important parts of a speaker recognition system are the feature extraction and the classification method. The aim of the feature extraction step is to strip unnecessary information from the sensor data and convert the properties of the signal, which are important for the pattern recognition task to a format that simplifies the distinction of the classes (Graevenitz, 2000). Usually, the feature extraction process reduces the dimension of the data in order to avoid the curse of dimensionality. The goal of the classification step is to estimate the general extension of the classes within feature space from a training set.

Feature extraction: Before identifying any voices or training a person to be identified by the system, the voice signal must be processed to extract important characteristics of speech. By using only the important speech characteristics, the amount of data used for comparisons is greatly reduced and thus, less computation and less time is needed for comparisons. The steps used in feature extraction are pre-emphasis, frame blocking, windowing, autocorrelation analysis, LPC analysis, campestral analysis and parameter weighting (Fig. 2).

Pre-emphasis: In this step, the signal is passed through a low-order FIR filter to spectrally flatten the signal and to make it less susceptible to finite precision effects. The transfer function of this filter is:

$$H(z) = 1 - az^{-1} \tag{1}$$

The value for a usually ranges from 0.9-1.0. For our system, we chose a = 0.9375.

Frame blocking and windowing: The signal is then put into frames (each 256 samples long). This corresponds to about 23 m of sound per frame. Each frame is then put through a Hamming window. Windowing is used to minimize the discontinuities at the beginning and end of each frame. The Hamming window has the form:

$$W(n) = 0.54 - 0.46*\mathrm{COS}\left(\frac{2*\pi*n}{N-1}\right);\ 0 \le \Pi \le N-1 \tag{2}$$

Where N is the number of samples per frame. In our system N = 256.

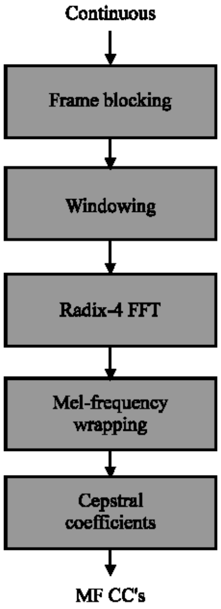


Fig. 2: Feature extraction processes

Autocorrelation analysis: In the third step, each windowed frame is auto correlated. This is done to minimize the mean square estimation error done in the LPC step. The equation for autocorrelation is:

$$rl(m) = \sum_{n=0}^{N-1-m} xl\,(n)xl(m+m);\ m = 0,1,...,p \tag{3}$$

Where p is the order of the LPC coefficients p = 13 in our system.

Linear Predictive Coding (LPC): The features for each frame are extracted by Linear Predictive Coding (LPC) which is represented by the following equation:

$$\hat{s}_n = \sum_{i=1}^{p} s_{n-1} \bullet a_i \tag{4}$$

Where Sn is the nth speech sample and the Ai are the predictor coefficients and Ŝn is the prediction of the nth value of the speech signal.

This is a finite-order all-pole transfer function whose coefficients accurately indicate the instantaneous configuration of the vocal tract.

Cepstral analysis: Cepstral coefficients are obtained from the LPC coefficients. Cepstral coefficients are used since they are known to be more robust and reliable than LPC coefficients, PARCOR coefficients, or the log area ratio coefficients. Recursion is used on this equation to attain the Cepstral coefficients

$$c(i) = a_i + \sum_{k=1}^{i-1}((1-\frac{k}{i}) \bullet a_k c_{i-k}); 1 < i \le p \qquad (5)$$

Here p = 13, a and k are the LPC coefficients and c (i) are the cepstral coefficients.

Parameter weighting: The cepstral coefficients are then passed through a parameter-weighting step to minimize their sensitivities. Low-order Cepstral coefficients are sensitive to the overall spectral slope and high-order Cepstral coefficients are sensitive to noise and other forms of noise-like variability. The weighted Cepstral coefficients are in the form:

$$\hat{c}_m = w_m c_m ; 1 \le m \le p \qquad (6)$$

Where W m is given by the following equation:

$$w_m = 1 + \frac{p}{2}\sin\left(\frac{\pi m}{p}\right); 1 \le m \le p \qquad (7)$$

Again p = 13 for our system.

The result is an i × j matrix, where i is the order of the LPC and j is the number of frames. Now that the important characteristics are extracted, the results can be used by VQ and DTW to compare the speaker and word, respectively.

Dynamic time warping: Dynamic Time Warping (DTW) is used for word recognition in our system. This is found to be a fast and easy way to recognize a spoken word.

Algorithm: The first step is to put 2 feature sets into a frame. One feature set is the unknown word and the other is a known word that is stored in the database.

Usually the known word is put on the y-axis and the unknown word is put on the x-axis of the frame. http://www.speaker-recognition.org/navAlg.html

First, the local distance is taken of the frame. This is the sum of Euclidean distances between both samples at certain points in time. Next, the accumulated distance is calculated to determine if the word matches. The equation to find the accumulated distance is given as:

$$D_A(1,1) = d(1,1) \bullet m(1)$$
$$D_A(i_x,i_y) = \min[D_A(i'_x,i'_y) + \xi((i'_x,i'_y),(i_x,i_y))(i'_x,i'_y)$$
$$d(X,Y) = \frac{D_A(T_x,T_y)}{M_\phi}$$

Where ξ is the weighted accumulated distortion (local distance) between point (i'_x, i'_y) and (i_x, i_y).

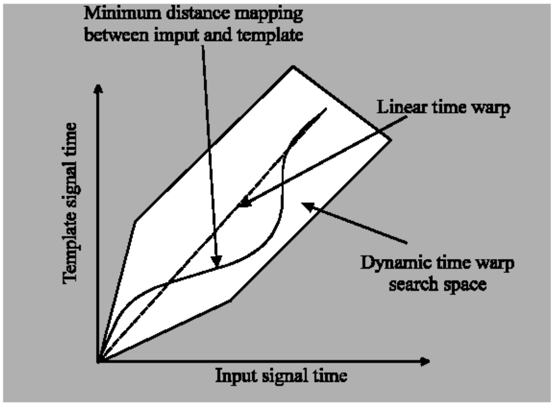

Minimum distance mapping between imput and template
Linear time warp
Template signal time
Dynamic time warp search space
Input signal time

Fig. 3: Dynamic time warping diagram

Three accumulated distances are calculated for the frame, the upper path, the middle path and the lower path. Figure 3 shows what each path means.

The gray area is all of the space where a distance can be calculated. The upper path is the distance near the gray boundary above the dashed line and the lower path is the distance near the gray boundary below the dashed line. The middle path is the distance near the dashed line and in the figure is represented by the solid line. The minimum distance of these three paths is taken as the final distance between these two samples.

Classification: Concerning the choice of the classification method, the kind of application of the speaker recognition system is crucial. For text independent recognition, speaker specific vector quantization codebooks or the more advanced Gaussian mixture models are used most often. For text dependent recognition, dynamic time warping or hidden Markov models are appropriate (Sadaoki, 2002).

Text independent recognition: Vector quantization is a technique, which is also used for speech coding. The training material is used to estimate a codebook. This includes mean vectors of feature vector clusters, which are given indices in order to identify them. For compression of speech, the index number of the nearest cluster is used instead of the original feature vector. In order to be able to reconstruct the original signal, a revertible feature computation method has to be chosen (i.e., the MFCC features described above cannot be used for speech coding). The quantization error in feature space is the mean distance between original feature vectors and nearest mean vectors (i.e., the feature used for reconstruction). Obviously, the quantization error depends on the similarity between training material used for estimation of the codebook and the audio signal that is compressed. For example, if a codebook is trained using speech signals, the compression of music with this codebook will result in a poor reconstruction for a listener as well as concerning the quantization error. This observation is also true concerning speaker specific codebooks, which are used for speaker recognition. The training material of a speaker is used to estimate a codebook, which is the model for that speaker. The classification of unknown test signals is based on the quantization error. For example, for an identification decision, the error of the test feature vector sequence about all codebooks are computed. The winner is the speaker which codebook has the smallest error between the test vectors and the corresponding nearest codebook vector. Gaussian Mixture Models (GMM) is similar to codebooks in the regard that clusters in feature space are estimated as well. In addition to the mean vectors, the covariance of the clusters are computed, resulting in a more detailed speaker model if there is a sufficient amount of training speech.

Text dependent recognition: Dynamic Time Warping (DTW) stores the labeled training vector sequences without any further processing. A test vector sequence is aligned to each of the training sequences such that a certain distance measure is minimized. Therefore, the classification algorithm can handle variations about the length of the phonemes an utterance consists of. Finally, a Hidden Markov Model (HMM) is a statistical model, which may be used for text dependent recognition of speakers. Roughly speaking, they can be viewed as a combination of the DTW and the GMM approach. A HMM has a number of states which model distinct parts of, for example, a user's password for a pass-phrase authentication system. The feature vectors, which are observed for the appropriate part of the pass phrase in training, are used to estimate a density function, e.g., a GMM. This is called the output density of the HMM state. A hidden Markov model is a more advanced representation for the pass phrase of a certain speaker, as the characteristic features for the phonemes that are present in the utterances are modeled statistically. Nevertheless, the DTW approach may be a better choice for a real-world speaker recognition system if the amount of available training data is not sufficient in order to reliably estimate the HMM's output densities.

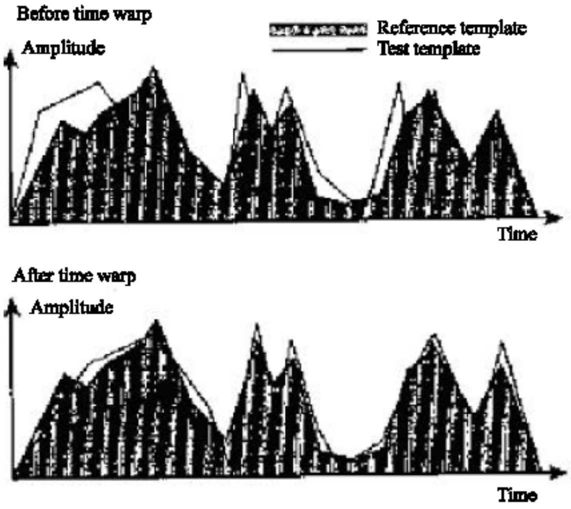


Fig. 4: An example of a template before and after dynamic time warping

Dynamic Time Warping (DTW) based classification: A classical approach to automatic speaker verification and recognition in the text dependent mode is based on pattern matching using spectral templates or spectrogram approach. In general, the speech signal is represented in terms of a sequence of feature vectors that characterize the behavior of the speech signal for a particular speaker. This time ordered set of features is called a template. A template can represent a multi-word utterance, a single word, a syllable or a phoneme.

In template-matching schemes, a comparison is made between an input utterance template and the reference template to verify the identity of the speaker. An important ingredient in these schemes is the need to normalize trial-to-trial timing variations. Note that these methods are only used in text-dependent systems. The normalization can be achieved by the Dynamic Time Warping (DTW) method. The DTW technique uses an optimum time expansion/compression function for producing non-linear time alignment. Figure 4 shows the templates before and after typical DTW. Note how the warping of the test template has improved the match between the two.

In Fig. 4, the speech frames that make up the test and reference templates are shown as scalar amplitude values plotted on a graph with time as the horizontal axis. A distance metric defined as a function of time, is computed between the two feature sets representing the speech data. A decision function can thus be computed by integrating the metric over time. In practice, the templates are multi-dimensional vectors and the distance between them is usually taken as the Euclidean distance. Another type of distance that is used to compare two sets of linear predictor coefficients is the Itakura distance.

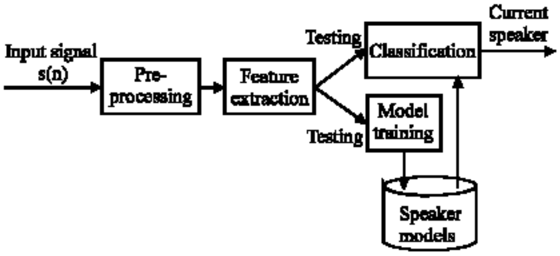


Fig. 5: The data flow diagram of a speaker identification system

RESULTS AND DISCUSSION

Program modules: The speaker identification program has two modules namely:

- User manager module
- Speaker identification module

The interfaces for each module are explained fully in the following sections.

User manager module: This module is the main segment that registers and trains user's voice patterns for future access to the computer system and only the administrator has access to this segment of the program.

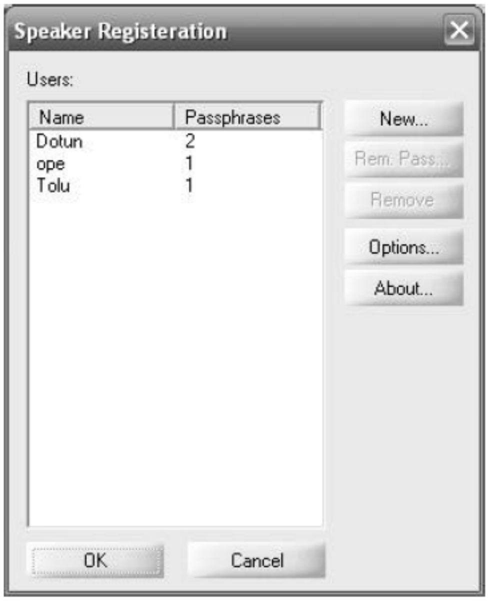
Speaker registration interface: The main interface for this module is the speaker registration interface. This interface as shown in Fig. 6 above displays the registered users and the number of pass phrases that each individual has registered on the system. It serves as a launching pad for adding new users and adding, modifying and deleting of pass phrases for existing users of the system.

New user interface: The dialog box shown in Fig. 7 gives the administrator the controls for adding new users or adding new pass phrases. It consists of a number of repetitive steps, which depend on how many repetitions the administrator configures the software to work on from the options dialog box.

The dialog box shown in Fig. 8 displays the speech waveform of the user during the enrolment process. The user proceeds to the next step only if the recorded pass phrases match (Fig. 9).

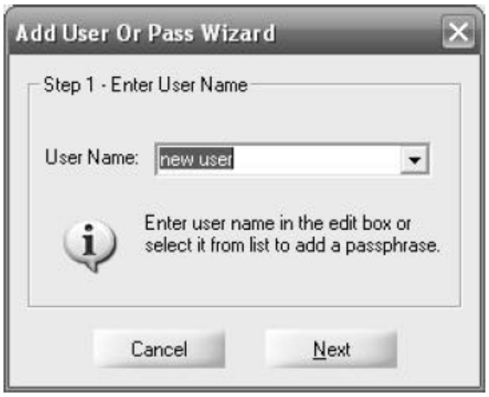
Options interface: The options dialog box (Fig. 10) allows the administrator to adjust the following parameters of the software:

- Number of Pass phrase repetitions: This is set between 2 and 6.



Name	Passphrases
Dotun	2
ope	1
Tolu	1

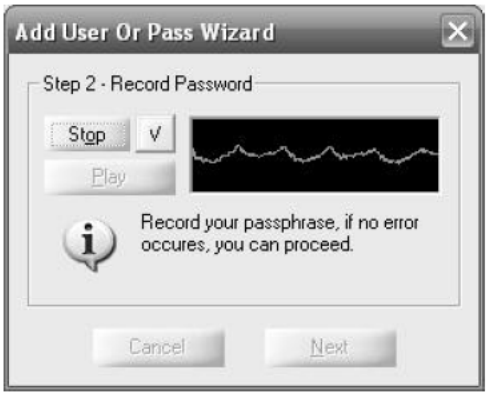
Fig. 6: Speaker registration interface



User Name: new user

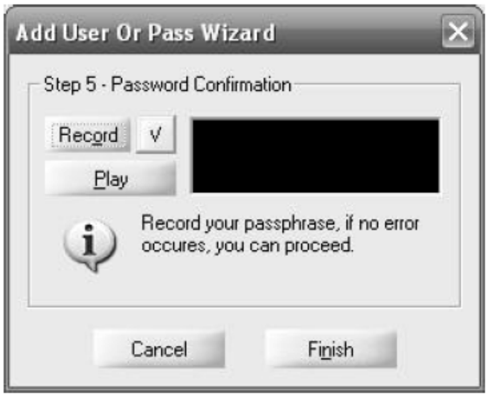
Enter user name in the edit box or select it from list to add a passphrase.

Fig. 7: New user interface



Record your passphrase, if no error occures, you can proceed.

Fig. 8: Pass phrase recording

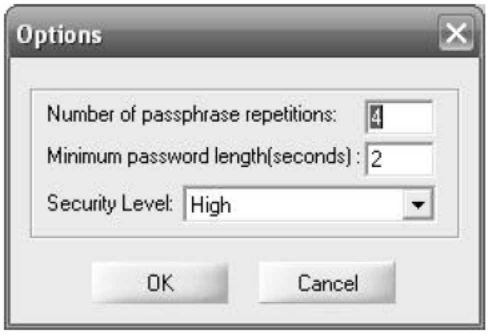


Add User Or Pass Wizard

Step 5 - Password Confirmation

Record your passphrase, if no error occures, you can proceed.

Fig. 9: Pass phrase recording confirmation



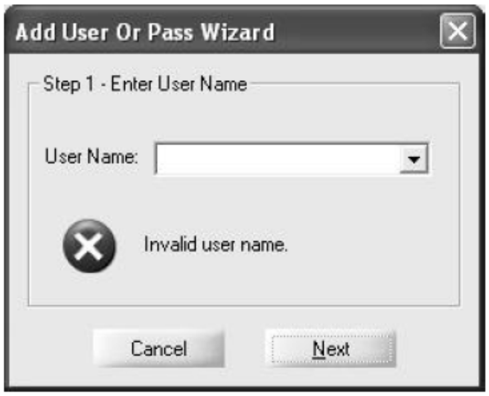
Options

Number of passphrase repetitions: 4

Minimum password length(seconds) : 2

Security Level: High

Fig. 10: Options



Add User Or Pass Wizard

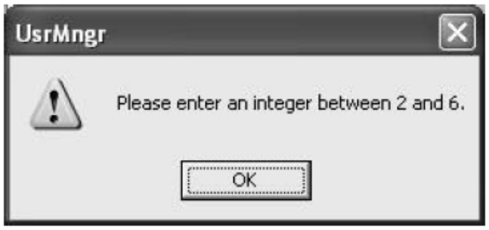
Step 1 - Enter User Name

User Name:

Invalid user name.

Fig. 11: Blank user name error

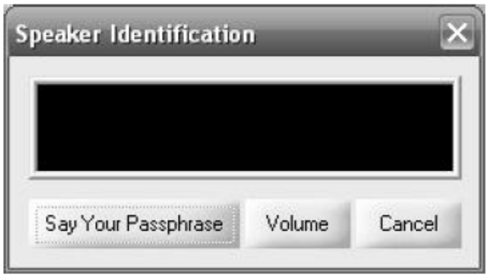
- Minimum password length (seconds): Which is also set between 2 and 6.
- Security Level: This determines the threshold levels of tolerance of the security system. The higher it goes, the more difficult to adapt to and vice versa. The security values are very high, high, medium, low and very low.



UsrMngr

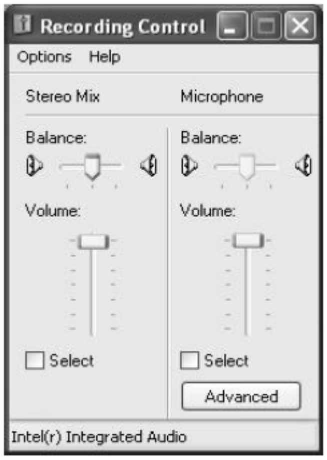
Please enter an integer between 2 and 6.

Fig. 12: Invalid integer range error



Speaker Identification

Fig. 13: Speaker identification dialog

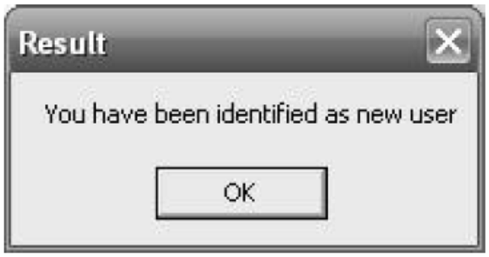


Recording Control

Fig. 14: Recording control dialog

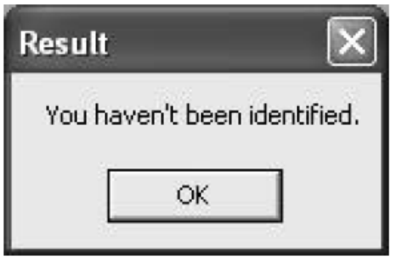
Error interfaces: These interfaces represent the software's reactions to wrong inputs from the administrator.

- Blank User Names: As shown in Fig. 11.
- Wrong Integer Values: This occurs when the administrator types in a pass phrase repetitions number or minimum pass phrase length that is less than 2 or greater than 6. This error was created to ensure efficient use of the system resources and is as shown in Fig. 12.



You have been identified as new user

Fig. 15: Successful login notification



You haven't been identified.

Fig. 16: Unsuccessful login notification

Speaker identification module: The speaker identification module (Fig. 13) is the part of the program that users use to get access to the computer system. Immediately Windows® has finished loading, the desktop is blocked and the user obtains access to the system only after a successful presentation of a pre-recognized pass phrase.

This dialog box also allows the user to adjust the system audio properties as shown in Fig. 14.

To signify a successful attempt, the system brings out the dialog box shown in Fig. 15. It acknowledges the successful login attempt and the recognized user name.

In the event of a failed attempt, the dialog box shown (Fig. 16) is displayed and it automatically displays the dialog box for logon.

CONCLUSION

On carrying out this project work, it can be concluded that unauthorized access to personal and corporate information on computer systems can be highly reduced by implementing speech-based biometric access control schemes. The simple and available equipments that are required buttress this claim: Typically, a sound card, microphone and a relatively quiet environment are the major requirements, which come along with the average computer system.

REFERENCES

Becchetti, C. and L.P. Ricotti, 1999. Speech Recognition, England John Wiley.

Graevenitz, G.A., 2000. Introduction to Speaker Recognition Technology, Germany, Bergdata Biometrics.

Jason Melby, 2003. Voice Biometrics. April 8, 2005 www.loop-start.com/Voice % 20Biometrics.pdf

Markowitz, J., 2002. Speech Recognition and Speaker Biometrics. August 4, 2005 http:// www.jmarkowitz.com/glossary.html#anchor_ glosstop.

Moshe Yudkowsky, 2002. Voice Biometrics and Application Security. Dr. Dobb's J., pp: 16-22.

Sadaoki Furui, 2002. Speaker Recognition. July 12, 2005, http://cslu.cse.ogi.edu/HLTsurvey/ch1node9.html.

The Speaker Recognition Homepage, 2002. Speaker Recognition-Algorithms. June 16, 2005 from http:// www.speaker-recognition.org/navAlg.html.