Speech Authentication based on Audio Watermarking

S.Saraswathi
Assistant Professor, Department of Information Technology, Pondicherry Engineering College, Pondicherry, India
swathimuk@yahoo.com

Abstract

With the rapid advancement of digital storage devices, speech recording has been widely used as evidence. This makes the speech authentication task more vulnerable. The reliability of stored speech requires proper security by the inhibition of access to content. Encryption is one way of preserving the digital information; however, digital forgery in signal processing level cannot be protected completely by this method. Digital forgeries of the signals are difficult to detect. One of the data hiding and extracting techniques is digital watermarking. This paper discusses on a method to perform speech authentication by means of watermarking technique. Watermark is embedded in the low intensity points detected in the speech signal. Rather sending the speech signal the extracted features are sent to the receiver for authentication. It is a blind watermarking technique in which the host signal is not required for watermark extraction. Authentication is done by detecting the errors in the signal based on their extracted features.

Keyword: Speech authentication, watermarking, Noise removal, feature extraction.

I. Introduction

In recent years, the distribution of digital media has grown rapidly due to the wide spread of the Internet. The copyright of these digital media is more difficult to manage. As a result, a technique called digital watermarking is introduced to protect the ownership of these contents. Digital watermarking can be realized by many different methods. In common to all of those methods, digital watermarks are embedded into the media contents as secret copyright identification code. A watermark hidden in a file cannot be detected by general user because the watermark will not deteriorate the quality of the file. Watermarks can be embedded into image, audio and video files. In Audio watermarking watermarks are introduced to audio files. In comparison to watermarking of other media, audio watermarking is relatively a recent subject.

Digital watermarking [1] is a technique of embedding information into a signal. The host signal that carries the watermark is called a cover signal. When the cover signal is an audio signal, the embedding technique is called audio watermarking. There are various purposes for audio watermarking. The original intention of watermarking is for copyright protection. Therefore, the most obvious purposes are the need for proof of ownership and the enforcement of usage policy. In addition, watermarking can also be used for fingerprinting and for adding additional features to a media. For a scheme to fulfill the purposes of watermark, a number of requirements have to be satisfied. The most significant requirements are perceptibility, reliability, capacity and speed performance.

Perceptibility: The most important requirement is that the quality of the original signal has to be retained after the introduction of watermark. A watermark should not be detected by listeners.

Reliability: Reliability involves the robustness and detection rate of the watermark. Watermarks have to be robust against intentional and unintentional attacks. The detection rate of watermark should be perfect whether the watermarked signal has been attacked or not. Otherwise, the watermark extracted is not useful for proof of ownership.

Capacity: The amount of information that can be embedded into a signal is also an important issue. A user should be able to change the amount of information embedded to suit different applications. There is a trade-off situation in information capacity of a watermark and its quality. A signal will be degraded if more information is embedded.

Speed: Watermarking may be used in real-time applications, such as audio streaming. The watermark embedding and extracting processes have to be fast enough to suit these applications.
Watermarking schemes can be classified based on the information required in the watermark extraction process. In private watermarking, the original data and the watermark are needed to verify the presence of a watermark. A secret key used in embedding is also needed. In semi-private watermarking, the original secret key and watermarks are needed in order to identify a watermark. The original signal is not required. Public watermarking, also called blind watermarking, requires the secret key used in embedding the watermark to be extracted. Earlier audio watermarking methods [2] employed the simplest way of, 'replacing', to embed data into host signal. One popular method was replacing the least significant bit (LSB) of each sample according to the watermark represented in a coded binary string. Moreover, it is more reasonable to modify LSB of frequency components, since the cochlea acts as the frequency filter bank and human beings are much more sensitive to low frequency part. Although high frequency components are perceptually insignificant, they are subjective to attacks as well, typically, lossy compression. Wavelet watermarking algorithm [3] employs wavelet decomposition/ reconstruction to realize watermark embedded in the host signal and to extract it from the watermarked signal. In the DC level shifting method [4], the watermark is embedded by shifting the DC level of the audio signal. Initially, an input signal is divided into frames of fixed length. Then, the DC level of each frame is calculated and the mean of the frames, is subtracted from the values in each frame. For this scheme of watermark embedding using Quadrature Mirror Filter (QMF) bank band division [5], the low frequency component of an input signal is extracted for watermark embedding. The QMF bank used in this scheme has a property of perfect reconstruction. A signal can be decomposed by the filter bank and recombined with no loss of data. A QMF bank contains levels of low pass and high pass filters. The filters used for dividing signals are called analysis filters. The frequency masking [6] technique exploits an effect called masking, where a faint sound is rendered inaudible by a louder sound that is played at the same time. In this case, the louder sound, also called the masker, is the host signal. The faint sound is the watermark. The masking operation is performed in the frequency domain.

The spread spectrum method [7] is a variation of the basic spread spectrum watermarking scheme. It improves the basic scheme by enhancing the robustness against de-synchronization attacks. In a basic spread spectrum scheme, the watermark embedded in a signal cannot be detected correctly if the signal is scaled along time or frequency axis. This form of attack is known as de-synchronization. The spread spectrum method overcomes this weakness by introducing an embedding technique called redundant encoding. A basic spread spectrum watermarking scheme [8] involves a transformation to be performed to an audio signal at first. The transformation converts a signal from time domain to frequency domain.

This paper discusses on the results obtained for a blind audio water marking technique. The paper is organized as follows: section 2 describes the proposed system and overall module of the project, section 3 describes the implementation details of each module of the project and section 4 analyses the results and future enhancements that can be done to the proposed work.

II. Proposed System

This project proposes a speech authentication by means of watermarking technique. Watermark is embedded in the low intensity points detected in the speech signal. Rather than sending the speech signal the extracted features are sent to the receiver for authentication. It is blind watermarking

technique where the host signal is not required for watermark extraction. Authentication is done by detecting the errors based on extracted features.

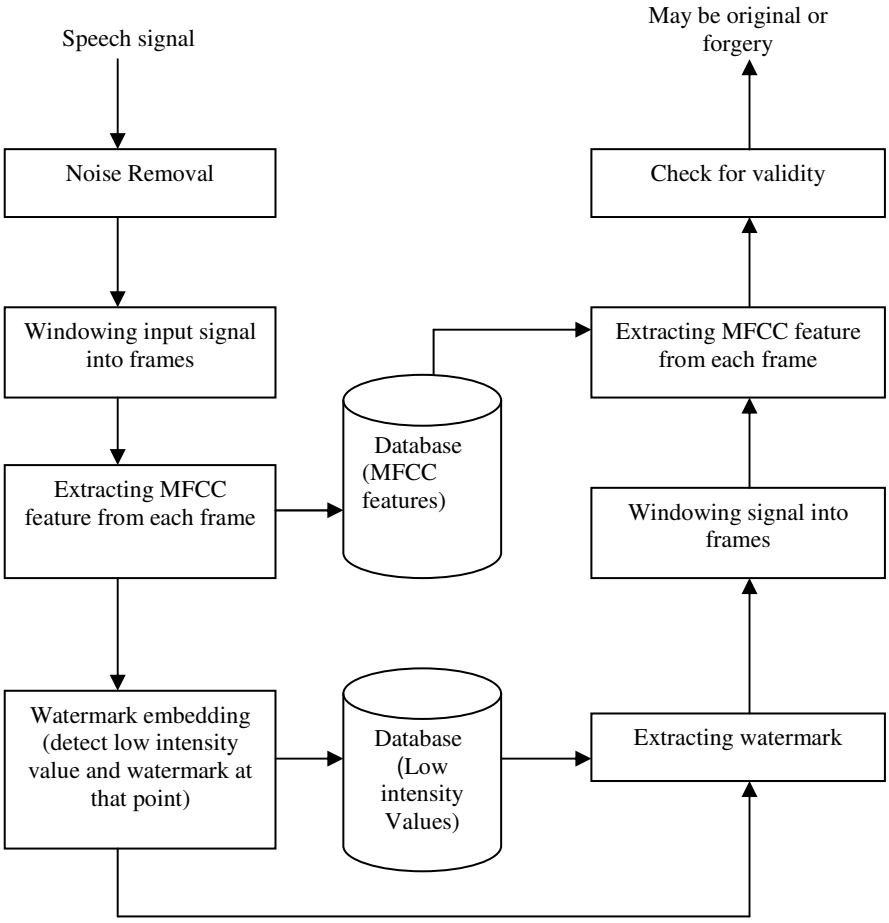

Figure 1. Overall Module of the proposed work

The overall block diagram of the proposed system is shown in figure 1. The speech signal is given as input to the system for authentication. The speech signal is divided into frames of fixed duration (60ms) and the Mel Frequency Cepstral Coefficients （MFCC) features of the frames are extracted. The extracted MFCC features are stored in the database. The low intensity points in the host speech signal are detected and stored in a database. The digital signal to be watermarked will contain information regarding the ownership of the speech signal. This signal is divided into frames. The duration of the frame depends on the size of the host speech signal. If the host signal contains *m* samples and the watermark signal contains *n* samples, then the watermark signal is divided into *m/n* frames of size n^2/m. These frames are inserted in the first *m/n* low intensity points of host speech signal. In this way the watermark is embedded in the host signal. The low intensity values are used for watermark extraction. The MFCC features extracted from the frames of the host signal is used for authentication. Watermark extraction, is done based on the low intensity values. The host signal is obtained and it is divided into frames of 60ms duration. The MFCC features extracted for each frame is compared with the features stored in the database for authentication. If there is any modification in the extracted features, forgery is detected. This work consists of the following modules:

A. Noise Removal

Noise removal is used to remove the background noise present in the speech signal. Power spectral subtraction method [9] is used for noise removal to adjust the subtraction factor. The adjustment is

according to the Signal to Noise Ratio (SNR). It calculates the number of initial silence present in the signal based on the shift percentage value. It calculates the noise length. Noise counter will return the number of previous noise frames present in the signal. It will detect the noise only on periods and will attenuate the signals. After detecting the noises they are removed and the noise free signal is reconstructed.

B. Feature Extraction

The speech signal is divided into frames of fixed duration (60ms) and the MFCC features [10] of the frames are extracted. The extracted features are used for authentication. MFCC (Mel Frequency Cepstral Coefficients) is one of the most widely used feature extraction technique. Since speech signal varies over time, it is more appropriate to analyze the signal in short time intervals where the signal is more stationary. To find the MFCC, the signal is split into short frames and a windowing function is applied for each frame to eliminate the effect of discontinuities at edges of the frames. The windowed signal is converted to frequency domain by taking the FFT (Fast Fourier transform) and Mel scale filter bank is applied to the resulting frames. After Mel frequency warping using the DFT function the signal is converted back to time domain. In most processing tools, it is not appropriate to consider a speech signal as a whole for conducting calculations. A speech signal is often separated into a number of segments called frames. This process of separation is known as framing. Fig. 2 illustrates how a signal might be divided into frames. Each frame will have the same number of samples although the last frame might have a small variance.

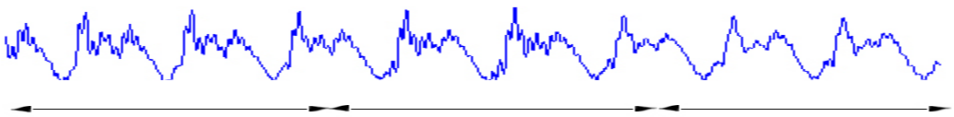

Figure 2. Signal divided into frames

There are many types of windowing that can be used in digital signal processing. These include Hamming, Hanning, Kaiser, Chebyshev etc. The purpose of windowing is to make the frame intense in some area while irrelevant in other areas. A typical window might be the Hamming window. The intensity of the window is much greater around the center than at the edges. When this window is multiplied point to point with a frame, the edges of the frame will become insignificant. Therefore, calculations on the frame will not be affected by the end data. Hamming windows have the property of low amplitudes at the edges in time domain.

The intension of using these properties is to compensate the effect of spectral leakage when a signal is divided into frames. The side lobes of these windows will break down the unwanted noise contributed in the signal. In the frequency representation, it is desirable to design the properties of the window to have a low noise bandwidth. This can be achieved by reducing the side lobe amplitudes. Different types of windowing functions are available, like Rectangular, Hamming, Barlett, Blackman, Kaiser, Bohman, Chebyshev, Hanning and Gausswin windows. Most of the windows have defined properties. However, the properties for Chebyshev and Kaiser Windows were manually defaulted to 60dB and 42dB of second side lobe attenuation.
Typically, hamming window is used for windowing,

$$w(n) = 0.54 - 0.46\cos\left(\frac{2\pi n}{N-1}\right), 0 \le n \le N-1 \tag{1}$$

The windowing is done to avoid problems due to truncation of signal. When a data is framed and windowed, the data at the ends of the frame is much likely to be reduced to zero. This will represent loss in information. An approach to tackle this problem is to allow overlapping in the sections

between frames. Overlapping will allow adjacent frames to include portions of data in the current frame. The edges of the current frame will be included as the center data of adjacent frames. Typically, around 60% of overlapping is sufficient to embrace the lost information. The analysis of a speech signal represented in frequency domain is of a great use. The process of Fourier transform converts a discrete signal x[n] from time domain representation into a frequency domain representation $X[e^{jw}]$ by the equation:

$$X[e^{jw}] = \sum_{n=-\infty}^{\infty} x[n]e^{-jwn} \tag{2}$$

However, in short domain Fourier transform, we can assume the sequence of signal is periodic with period N, and thus the equation of the Fourier transform can be represented as:

$$X[k] = \sum_{N=0}^{N-1} x[n]e^{-j2\pi kn/N} \tag{3}$$

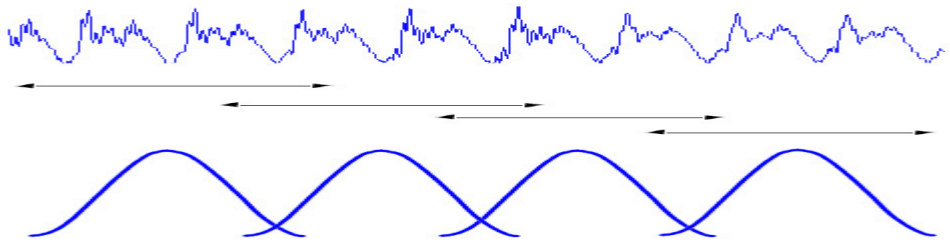

Figure 3. Overlapping of frames

The computation of X[k] is very inefficient and time consuming. The computation time is found to be proportional to (N^2). A set of computational algorithms known as the fast Fourier transforms (FFT) is used. The computation time was decreased dramatically to a proportion of $N\log_2 N$. The algorithm of FFT is based on the concept that the processing time of multiplication is longer than addition. Therefore, FFT is used to modify the calculation from a series of multiplications to a series of additions for faster computation.

The psychophysical studies have shown that human perception of the frequency contents of sounds for speech signals does not follow a linear scale. Thus for each tone with an actual frequency, f, measured in Hz, a subjective pitch is measured on a scale called the 'mel'scale. The mel frequency scale is a linear frequency spacing below 1000 Hz and a logarithmic spacing above 1000Hz.As a reference point, the pitch of a 1kHz tone, 40 dB above the perceptual hearing threshold, is defined as 1000 mels. Therefore we can use the following approximate formula to compute the mels for a given frequency f in Hz:

$$\text{Mel}(f) = 2595 * \log_{10}(1 + f/700) \tag{4}$$

One approach to simulating the subjective spectrum is to use a filter bank, spaced uniformly on the mel scale. The filter bank has a triangular band pass frequency response, and the bandwidth is determined by a constant mel frequency interval. The modified spectrum consists of the output power of these filters. The number of mel spectrum coefficients, k, is typically chosen as 20. Note that this filter bank is applied in the frequency domain; therefore it simply amounts to taking those triangle–shape windows on the spectrum. A useful way of thinking about this mel-wrapping filter bank is to view each filter as a histogram in the frequency domain. In this final step, the mel spectrum is converted back to time domain. The result is called the mel frequency cepstrum coefficients (MFCC). The cepstral representation of the speech spectrum provides a good

representation of the local spectral properties of the signal for the given frame analysis. Because the mel spectrum coefficients are real numbers, we can convert them to the time domain using discrete cosine transform (DCT).

C. Watermark Embedding

In this module the watermark is embedded in the speech signal. The digital signal to be watermarked will contain information regarding the ownership of the speech signal. This signal is divided into frames. The duration of the frame depends on the size of the host speech signal. If the size of host signal contains *m* samples and the size of the watermark signal is n, then the watermark signal is divided into *m/n* frames of size n^2/m. These frames are inserted in the first *m/n* low intensity points of host speech signal. The locations of low intensity points are stored in the database. These points are also used in watermark extraction process.

D. Watermark Extraction

The watermarked signal, the position of low intensity point's, *m* (the no of samples in host signal), *n* (the no. of samples in watermarked signal) and the MFCC features extracted for each frame in the host signal are passed to the receiver side. By using low intensity points the frames of the watermarked signal of size n^2/m are extracted from the host signal. After extraction the signal is divided into frames by using hamming window technique. From the frames of the host signal their mel frequency cepstral coefficient (MFCC) values are extracted.

E. Authentication

The authentication is a service concerned with the assurance that a communication is authentic. In this work authentication is performed based on the extracted features. The features received from sender are checked with the obtained features. If there is any modification in the coefficient value the errors are identified. If there is no modification in the coefficient value then it is an original signal.

III. Implementation details

This describes the implementation details of each module of the project. It consists of preprocessing, extracting features; watermark embedding, watermark extraction and authentication. MATLAB 7.0, a high-performance language for technical computing is used for the implementation of this work. It integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are expressed in familiar mathematical notation.

A. Preprocessing

This stage includes the noise removal and framing modules.

Noise Removal: Spectral subtraction technique is used to remove the background noise present in the input signal. The number of initial silence present in the signal based on the shift percentage value is calculated. It will chop the signal into frames based on the number of initial silence. It will detect the noise only on periods and attenuate the signal. Using FFT it will reconstruct the signal. A noise free output signal is generated.

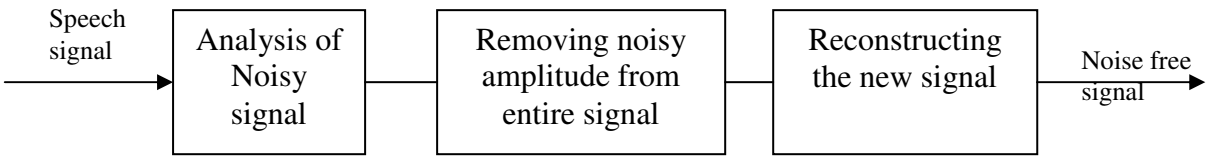

Figure 4. Spectral subtraction method

Framing: The next step in pre-processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and ending of each frame. By windowing we tend to minimize the spectral distortion as it tapers the signal to zero at the beginning and end of each frame. Typically Hamming window is an efficient technique for framing the signal.

B. Feature Extraction

To find the MFCC, the signal is divided into short frames and a windowing function is applied for each frame to eliminate the effect of discontinuities at edges of the frames. The windowed signal is converted to frequency domain by taking FFT and Mel scale filter bank is applied to the resulting frames. After Mel frequency warping logarithm of the framed signal is passed to the inverse DFT function for converting the signal back to time domain. As a result of the final step, 13 coefficients named MFCC for each frame are obtained. The MFCC coefficient values calculated for the frames are stored in the database for feature comparison.

C. Watermark Embedding

The low intensity points in the host signal are detected. The number of low intensity points (m/n points) depends upon the ratio of the host signal and watermarked signal. The position of the low intensity points is stored in the database. The watermark signal is divided into (m/n) frames of size n^2/m. Finally watermark embedding is done on the low intensity points of the host signal. The MFCC features of the host signal, position of low intensity points, size of watermark frame embedded in each point are sent to the receiver side for extraction.

D. Watermark Extraction

By using low intensity points the watermarked signal is extracted from the host signal. After extraction the host signal is divided into frames by using hamming window technique. Finally we extract the features from signal by using mel frequency cepstral coefficient (MFCC). These features are used for authentication purpose.

E. Authentication

Authentication is performed based upon the extracted features. The feature comparison between received and obtained feature is performed by Euclidean metric. The Euclidean distance or Euclidean metric is the "ordinary" distance between two points that one would measure with a ruler, which can be proven by repeated application of the Pythagorean Theorem. By using this formula as distance, Euclidean space becomes a metric space. The Euclidean distance between points $P = (p_1, p_2, \ldots, p_n)$ and $Q = (q_1, q_2, \ldots, q_n)$, in Euclidean *n*-space, is defined as:

$$\sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_n - q_n)^2} = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}. \qquad (5)$$

Euclidean distance between MFCC features of original signal with MFCC features of received signal is calculated. If the result is zero, then the signal is authenticated as original one.

IV. Results and Discussion

Speech signal is input to the system. The data were recorded using an 8 kHz sampling rate and a 16 bit resolution and the length was 60s. The recording has been done in in-door environments using quiet and noisy office. The embedding time was 9% of the playing time. Preprocessing is done to remove the noise from input signal. After watermarking different types of errors like insertion, deletion and replacement are introduced to check the performance.

Insertion: Signals of different durations were inserted to the watermarked signals at various locations like beginning, middle and ending. The performance analysis after inserting, signals of

different durations in the beginning, middle and the ending of the watermarked signal was evaluated**.** The results obtained are depicted in table 1.

Table 1: Performance analysis after inserting, signals of different durations in the beginning, middle and the ending of the watermarked signal

Size of the inserted signal	**Position of insertion**	**Percentage of similarity between send and received signals (based on MFCC)**
10ms	Beginning	20
	Middle	45
	Ending	90
20ms	Beginning	13
	Middle	34
	Ending	86
30ms	Beginning	9
	Middle	26
	Ending	81
40ms	Beginning	5
	Middle	22
	Ending	75
50ms	Beginning	0
	Middle	10
	Ending	68

The average percentage of similarity of the signal with the original signal was 9.4% in the case of insertion at the beginning, 27.4% for insertion in the middle and 80% for insertion at the end. The insertion at the beginning will tend to change the overall characteristics features of the signal. Since the MFCC features are checked for every frames of size 60ms, the checking was done only on the signals with duration less than 60ms. Insertion in the middle will change partial characteristics of the signal, so its performance quiet better than insertion at the beginning. The insertion at the end will only tend to change the characteristics of the last frame of the signal. Hence the percentage of similarity is more in this case.

Deletion: Signals of different durations were deleted at the beginning, middle and ending of the watermarked signals and their percentage of similarity with the original signal was evaluated. The results obtained are depicted in table 2. The average percentage of similarity of the signal with the original signal was 5.8% in the case of deletion at the beginning, 26.4% for deletion in the middle and 88.6% for deletion at the end. The deletion in the beginning will tend to change the overall characteristics features of the frames. Deletion in the middle will change characteristics of the frames obtained from the middle part of the signal. Its performance is quiet better than deletion at the beginning. The deletion at the end will only tend to change the characteristics of the last frame of the signal. Hence the percentage of similarity is more in this case.

Table 2. Performance analysis after deleting signals of different durations in the beginning, middle and the ending of the watermarked signal

Size of the deleted signal	**Position of deletion**	**Percentage of similarity between send and received signals (based on MFCC)**
10ms	Beginning	10
	Middle	38
	Ending	95
20ms	Beginning	8
	Middle	32
	Ending	93
30ms	Beginning	6
	Middle	26
	Ending	90
40ms	Beginning	5
	Middle	20
	Ending	85
50ms	Beginning	0
	Middle	16
	Ending	80

Replacement: Signals of different durations were replaced at the beginning, middle and ending of the watermarked signals and their percentage of similarity with the original signal was evaluated. The performance analysis after replacing the signals of different durations in the watermarked signal is listed in table 3. The effect of similarity was the same for replacement at different locations. This is due to the fact that the replacement will only affect the characteristics features of one frame at an instant. When there was an increase in the size of the replaced signal the percentage of similarity with the original signal decreased drastically.

Table 3. Performance analysis after replacing signals of different durations in the watermarked signal.

Size of the replaced signal	**Percentage of similarity between send and received signals (based on MFCC)**
10ms	98
20ms	95
30ms	90
40ms	87
50ms	82

V. Conclusion

In this work previously available watermarking techniques have been studied and the drawbacks are identified. In the proposed system, watermarking is done in the low intensity points of the speech signal. Rather than sending the speech signal the extracted features are sent to the receiver for authentication. It is blind watermarking technique the host signal is not required for watermark extraction. Authentication is done by detecting the errors in the received signal based on the extracted features. Further enhancements that can be done to this work are: to identify the type of errors like insertion, deletion and substitution in the signal, to compare the similarity among the signals based on features other than MFCC and to improve the security enhanced speaker verification system based on speech signal watermarking.

References

[1] R.J Anderson, M.G Kuhn and F.A.P. Peticolas, "Information Hiding-A survey"- IEEE proceedings, volume 42, pp 675-683, July 1999

[2] LV Jiu-ming, Luo Jing-qing and Yuan Xue-hua, "Digital Watermark technique Based on Speech Signals" - IEEE proceedings, volume 11, pp 215-225, 2004

[3] M.S. Hsieh, D.C. Tseng Y.H. Huang, " Hiding Digital watermarks using Wavelet Transform, IEEE Transactions on Industrial Electronics, volume 48, pp 875-882,oct 2001

[4] Levent M. Arslan and Umut Uludag "Audio watermarking using DC level shifting" Advanced topics in speech processing using DC level shifting, Jan 2001

[5] Saito, S., Furukawa, T. and Konishi, K. "A digital watermarking for audio data using band division based on QMF bank." IEEE International Conference on Acoustics, Speech, and Signal Processing, volume 4, pp 3473-3476, 2000

[6] Nedeljko Cvejicand and Tapio Seppanen "Spread-spectrum audio watermarking using frequency hopping and attack characterization" –Elsevier science publishers, volume 84, pp 207-213,2003

[7] Qiang Cheng, Jeffrey, "Spread spectrum signaling for speech watermarking" IEEE International Conference on Acoustics, speech and signal processing, volume3, pp 1337-1340,May 2001.

[8] Darko Kirovski and Henrique S. Malvar, Fellow "Spread-Spectrum Watermarking of Audio Signals" IEEE transactions on Signal processing, volume 51, pp 1020-1033,2003

[9] S.F.Boll, "Suppression of acoustic noise in speech using spectral subtraction"-IEEE transactions on Acoustics, Speech and signal processing, 2001 volume 27, pp 113-120,2000

[10] D.Chasan, "Speech reconstruction from mel frequency cepstral coefficients and pitch"- IEEE transactions on Signal processing, 2000

Dr. S.SARASWATHI, is an Assistant Professor, in the Department of Information Technology, Pondicherry Engineering College, Pondicherry, India. She was born on 29th September 1971 at Karaikal, Pondicherry, India. She completed her B.Tech. degree in Computer Science and Engineering, in the year of 1993, from Pondicherry Engineering College, Pondicherry, India. She has completed her M.Tech. in Computer Science and Engineering, in the year 1995, from Pondicherry University, Pondicherry. She has completed her Ph.D. in the area of Speech Processing at Anna University, Chennai, India, in the year 2008. She has been continuing in the academic streamline from 1995. She has so far published 7 papers in International Journals and 20 papers in International conferences. Her area of interest includes Artificial Intelligence and Expert System, Speech Processing, Natural Language Processing, Agent-based Computing and Database Management System.