International Journal of Advanced Information Science and Technology (IJAIST) ISSN: 2319:2682
Vol.1, No.6, October 2012 DOI:10.15693/ijaist/2012.v1i6.44-51

Approaches to Iterative Speech Feature Enhancement and Recognition using HMM and Modified HMM

Deny.J[1], Densi.J[2], Sivasankari.N[3]

[1]*Assistant Professor,ECE Depaement,PSN Enginnering College*
[2]*Assistant Professor,CSE Depaement,Asan Memorial College of Engg &Tech*
[3]*Assistant Professor,ECE Depaement,Kalasalingam University*

Abstract :

In speech recognition, Hidden Markov Models (HMMs) are commonly used for speech decoding. Initially, n numbers of speech signals of n number of speakers are stored in the database. Then the features of the real time input signal is extracted and is compared with that of the signals stored in database using the MFCC (Mel-Frequency Cepstral Coefficients) algorithm. The speech recognition is first tested in the absence of noisy environments and then it is tested in the presence of noisy environments. The algorithm used to test the speech signal authentication in the absence of noisy environments is performed using MFCC and HMM algorithms. The proposed algorithm is applicable for adaptively estimating and removing background noise from speech signals. Unlike noise reduction and speech enhancement algorithms currently available on the market, the new algorithm is capable of cleaning noisy speech even in severe noisy environments without any distortion to the speech signal. The proposed algorithm called Modified Hidden Markov Model (Modified HMM) and Modified Mel-Frequency Cepstral Coefficients (Modified MFCC) are capable of doing Speech signal enhancement and recognition and the Expectation Maximization (EM) algorithm is used to estimate the features of the enhanced speech signal. This project may be used in various application areas such as, for authentication number of employees working in offices, in military applications and also in all possible security applications.

Keywords: **MFCC, HMM, speech recognition, speech authentication, database.**

I INTRODUCTION

Speech processing is the study of speech signals and the processing methods of these signals. It is one of most important branches in digital signal processing The signals are usually processed in a digital representation, so speech processing can be regarded as a special case of digital signal processing, applied to speech signal.. Speech signals can be used for speech recognition, speaker recognition or voice command recognition systems [1]. Proposed voice command recognition system includes two main stages. First stage contains feature extraction and storage of extracted features as training data. Second stage is test. In this stage, features of a new entered command are extracted. These features are used in order to make comparison with stored features to recognize command.

1.1 Speech signal

It is the variation of pressure, from atmospheric pressure, as a function of time, caused by traveling waves from the speaker's mouth (apart from nostrils, cheeks and throat). The energy of speech during 1 second – 2 x 10 –5 Joules

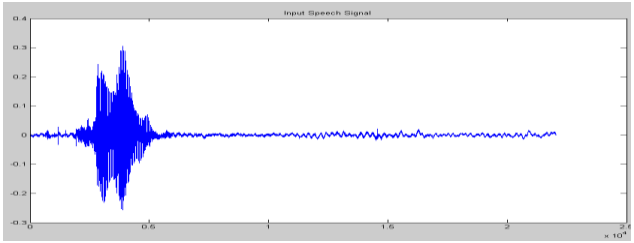

Fig 1 General illustration of a human speech signal

1.1 Speech recognition

The process of converting words spoken in a form that is machine readable is termed as speech recognition. The speech

recognition system has found many applications in today's techno-savvy world.. On the basis of vocabulary and the number of users, speech recognition systems are categorized into Small vocabulary/many-users and Large vocabulary/limited-users[5]. The speech recognition technology is also categorized as discrete speech recognition and continuous speech recognition systems. In the discrete speech recognition systems, the dictator has to take a pause after every word spoken. The continuous speech recognition system understands words that are spoken in a normal manner.

Fig 2 An Overview of speech recognition

1.2 Speaker recognition

Speaker recognition is the computing task of validating a user's claimed identity using characteristics extracted from their voices. It also refers to a group of technologies that use information extracted from a person's speech to perform operations such as Speaker Identification and Speaker Verification.

1.2.1 Categories of speaker recognition

Speaker recognition system falls into two categories.

They are

1 Text-Dependent

2 Text-Independent.

In a text-independent recognition, speaker model capture the characteristics of somebody's speech which show up irrespective of what one is saying. The recognition system does not have any information about the content of training and test utterences.

A text dependent system relies on the restriction that the text that is said in training is identical to the test utterance. The recognition of the speaker's identity is based on his or her speaking one or more specific phrases, like passwords, card numbers, etc

1.2.2 Phases of speaker recognition

There are two phases of speaker recognition. They are

1 Enrollment phase.

2 Verification phase.

The Enrollment phase is also known as training phase. In this phase, speaker's voice is recorded and typically a number of features are extracted to form a voice print, template, or model. In case of speaker verification systems, in addition, a speaker-specific threshold is also computed from the training samples.

The Verification phase is also known as testing phase. During this phase, the input speech is matched with stored reference model and recognition decision is made.

1.2.3 Operations of speaker recognition

In the speaker recognition the information extracted from a person's speech to perform operations such as

1 Speaker identification

2 Speaker verification.

In forensic applications, it is common to first perform a speaker identification process to create a list of "best matches" and then perform a series of verification processes to determine a conclusive match.

Identification is the task of determining an unknown speaker's identity. A speaker identification system gets a test utterance as input. The task of the system is to find out which of the training speakers made the test utterance. So, the output of the system is the name of the training speaker, or possibly a rejection if the utterance has been made by an unknown person. Speaker identification is a 1:N match where the voice is compared against N templates.

Speaker identification systems can also be implemented covertly without the user's knowledge to identify talkers in a discussion, alert automated systems of speaker changes, check if a user is already enrolled in a system, etc.

II THE SPEECH SIGNAL RECOGNITION PROCESS

The speech signal is compared with that of the speech signals already stored in the database in the absence of noisy environments. This includes two modules

1 Feature extraction module.

2 Speech signal recognition module.

In the feature extraction module, the features of n number of speech signals of various speakers are stored in the database.

In the speech signal recognition module, the speech signal which is given as the input is compared with those of the signals stored in the database and is authenticated.

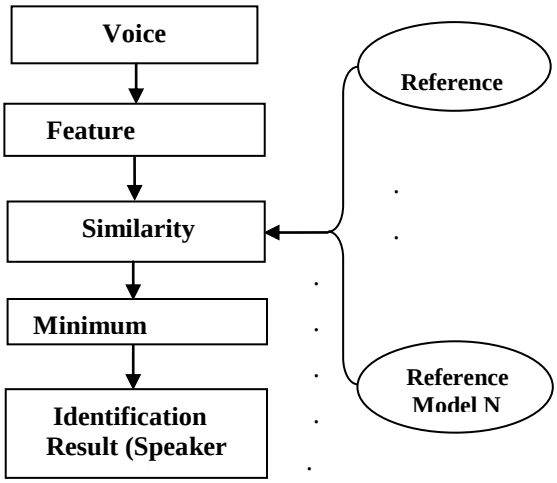

Fig 3 Flow diagram for speaker recognition

2.1 Feature Extraction and Parameter Estimation

The data type of the speech signal is stored in the form of wave file. The system only accepts wave file format. The voice record should be taken in a quiet environment. To store the speech signal in the database, various features of that signal are extracted. The parameters calculated are

1 Cepstral co-effecients represented in mel scale.

2 Linear Prediction Coefficients

3 Perceptual Linear Prediction Coefficients

4 Spectral Parameters

2.1.1 Cepstral co-effecients represented in mel scale

MFCC Algorithm is used in order to obtain the cepstral representation of the speech signal. This is one of the parameter extracted from the speech signal[2].

speech

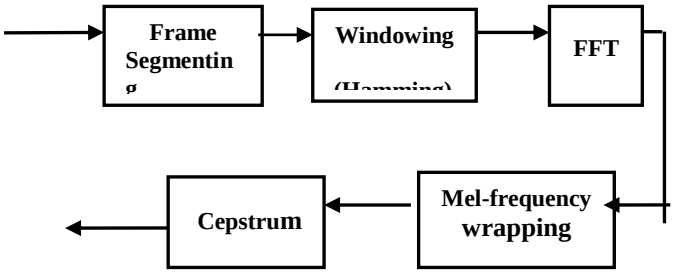

Fig 4 Block diagram of MFCC algorithm

- *Frame Segmenting:* The continuous speech signal is blocked into frames of *N* samples, with adjacent frames being separated by *M* (*M* < *N*). The first frame consists of the first *N* samples.[3] The second frame begins *M* samples after the first frame, and overlaps it by *N* - *M* samples. Similarly, the third frame begins 2*M* samples after the first frame (or *M* samples after the second frame) and overlaps it by *N* - 2*M* samples.
- *Windowing:* The next step in the processing is to window each individual frame so as to minimize the signal discontinuities at the beginning and end of each frame[7].The concept here is to minimize the spectral distortion by using the window to taper the signal to zero at the beginning and end of each frame.
- *Calculating FFT:* This converts each frame of *N* samples from the time domain into the frequency domain.
- *Mel-frequency wrapping:* The input signal uses the mel filter bank, spaced uniformly on the mel scale. That filter bank has a triangular bandpass frequency response, and the spacing as well as the bandwidth is determined by a constant mel frequency interval[1]. This filter bank is applied in the frequency domain, therefore it simply amounts to taking those triangle-shape windows in on the spectrum.

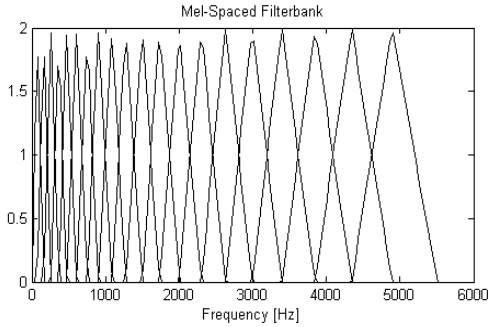

Fig 5 Illustration of mel filter bank

- *Cepstrum computation:* The signal obtained in the form of log Mel spectrum is converted back from frequency domain to time domain using Discrete Cosine Transform (DCT)[1].

2.1.2 Linear prediction co-efficients

Linear prediction co-effecients determines the coefficients of a forward linear predictor by minimizing the prediction error in the least squares sense. This involves process of circular buffering, initializing the arrays for buffer generation.Then the data is then rounded off towards infinity and minus infinity and hence the downsampling is performed. Then the minimum value from the downsampled output is obtained. Convolution is performed and then multiplied with the window function.

2.1.3 Perceptual linear prediction co-effecients

This is another feature extracted in the speech signal. In this hamming window is used. In this feature, toeplitz matrix is used. A Toeplitz matrix is defined by one row and one column. A symmetric Toeplitz matrix is defined by just one row. It generates Toeplitz matrices given just the row or row and column description. Eg- T = toeplitz(c,r) returns a nonsymmetric Toeplitz matrix T having c as its first column and r as its first row.

2.1.4 Spectral parameters

Spectral estimation describes the distribution (over frequency) of the power contained in a signal, based on a finite set of data. Estimation of power spectra is useful in a variety of applications, including the detection of signals buried in wide-band noise. The various spectral parameters used to extract the features of the speech signal are

- Spectral descriptor
- Arithmetic Mean
- Geometric Mean
- Spectral Flatness
- Spectral Skewness
- Spectral Kurtosis
- Spectral Centroid

Spectral descriptor: The welch spectum is used to obtain the properties of the spectrum. It returns a default Welch spectrum object Hs, that defines the parameters for Welch's averaged, modified periodogram spectral estimation method. The object uses default values such as estimation method, window name, fft length, overall percent, segment length.

The values declared for estimating Hs is given below

Hs =

EstimationMethod: 'Welch'

FFTLength: 'NextPow2'

SegmentLength: 64

OverlapPercent: 50

WindowName: 'Hamming'

SamplingFlag: 'symmetric'

Then the power spectral density (psd) is calculated for the obtained Welch spectrum object.

The values declared are

Power Spectral Density =

Name: 'Power Spectral Density'

Data: [16385x1 double]

SpectrumType: 'Onesided'

Frequencies: [16385x1 double]

NormalizedFrequency: false

Fs: 11025

The final step involves the extraction of spectral descriptors from the power spectral density of the signal.

Arithmetic mean: The arithmetic mean of the data series is computed and returns it to the determined structure. The

arithmetic mean of the spectral descriptors of the signal is obtained.

Geometric mean: The geometric mean of a speech sample is calculated. For vectors, geomean(x) is the geometric mean of the elements in x. For matrices, geomean(X) is a row vector containing the geometric means of each column. The geometric mean is given by the formula

$$m = \left[\prod_{i=1}^{n} x_i\right]^{\frac{1}{n}}$$

,where i varies from 1 to n.

Spectral flatness: The spectral flatness is to improve the accuracy of the signal. It is obtained by dividing geometric mean by arithmetic mean.

SPECTRAL FLATNESS = GEOMETRIC MEAN / ARITHMETIC MEAN

Spectral skewness: Spectral skewness is another feature extracted from the speech signal. Skewness is a measure of the asymmetry of the data around the sample mean.

- If skewness is negative, the data are spread out more to the left of the mean than to the right.
- If skewness is positive, the data are spread out more to the right.

Spectral kurtosis: Kurtosis is a measure of the "peakedness" of the probability distribution of a real-valued random variable[4]. Spectral kurtosis is a representation of the kurtosis of each frequency component of a process.

Spectral centroid: The centroid is calculated by performing the cumulative sum of the spectral speech signal. Then the obtained output signal which is in the form of a matrix of data is sorted in descending order. After that the threshold value for the signal is obtained using the formula

Threshold value = 0.7 * max(Output signal)

where max(Output signal) returns the largest element in Output signal. After this, the threshold value is compared with the each of the data of the output signal which is in the form of matrix using a for loop. An array is initialized for incrementation which is assumed as spectral cross. If the data is greater than the threshold value, then the spectral cross incremented by one. The spectral cross is incremented by one for each comparison. Then a finite sequence of datas are taken from the sorted output and the datas and its locations are stored in two separate arrays.

2.2 Speech Signal Recognition and Authentication

The real time speech signal which is given as the input in the absence of noisy environments is compared with that of the speech signals stored in the database and is recognized and authenticated. This process is performed by HMM algorithm.

2.2.1 Hidden Markov Model – a Basic Description

The Hidden Markov Model is used for speech recognition, thus it is useful in particular for text-dependent speaker recognition[2]. HMM is a stochastic model. The HMM can be viewed as a finite state machine. Each state (node) in it has an associated probability density function (PDF) for the feature vector. On moving from one state to another the probability of that transition is defined. Only the first and the last states are not-emitting states, since the first is always where it is started and the last one is the one where we always end our transitions, i.e. there are no incoming transitions into the start state and there are no output transitions from the end state. Every emitting state has a set of outgoing transitions and the sum of the probabilities for those transitions is equal to one, since the transition from non-final state always must occur.

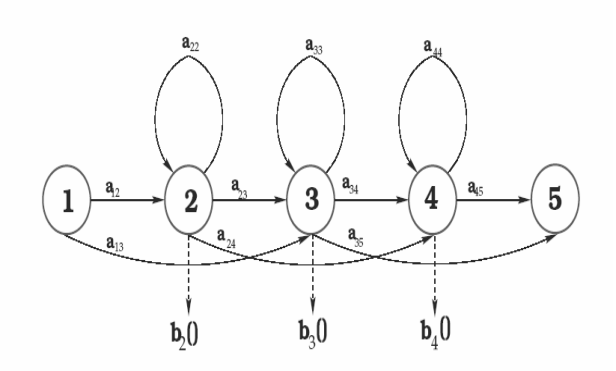

Fig 6 A simple description of HMM

In Fig 6, the numbers are the states. 1st is the start state and 5th is the end state. The b_i are the PDFs. The a_{ij} are the probabilities.

2.2.2 Feature comparison

In the Speaker Verification module, each Mel-Cepstral Co-effecient vector of the test speech is compared with the codebooks to calculate its distances (e.g., Euclidean distance) to each codebook[6]. The codebook vector closest to the test vector is found. The corresponding minimum Euclidean Distance, or Distortion Factor, is then stored until the Distortion Factor for each test vector has been calculated[6]. The Average Distortion Factor is then found and normalized.

2.2.3 The Speech Recognition and Authentication Process

The speech signal recognition module involves a series of steps.

- Apply the features to the signal to be compared.
- Compute the difference
- Compute the sum of trials.
- Perform mapping of the signal.
- Perform Mutation.
- Speech Verification Process

Apply the features to the signal to be compared: When the speech signal is given at real time, in the absence of noisy environments, the features of the speech signal such as cepstral co-effecients, linear prediction co-effecients, perceptual linear co-effecients are calculated. Then the file in which the signals stored in the database is opened. The end of file process is then verified. Then the minmax value for the output matrix is obtained. The minmax is the range of matrix rows which implies the minimum and maximum values for each row of the specified matrix.

Compute the difference: The speech signals which are stored in the database are called *reference signals or database signals*, which is in the form of reference matrix. The signal given for comparison with the signals stored is called *test signal*, which is stored in the form of test matrix. After applying the features to the signal given for verification, the difference between each of the database signal and test signal is computed. After applying the features to the signal given for verification, the difference between each of the database signal and test signal is computed.. Then the sum of this value is stored in the array containing the difference between the test signal and each of the database signals. The matrix element which returns the value zero is considered as the signal to be recognized. Example - If the 1st element of the matrix is zero then that is the speech signal which matches with the database. Hence the first signal is matched with the real time speech input given. This is a rough computation of the authentication process.

Compute the sum of trials: This involves the computation of cumulative matrix , trial matrix and the comparison of the cumulative and trial matrix. The cumulative sum of the output matrix is obtained. Then the repeat index is calculated by storing the row elements and column elements in separate arrays. The trials is calculated by multiplying the matrix of cumulative sum with that of the rand function which generates arrays of random numbers whose elements are uniformly distributed in the interval (0,1).The trial matrix is generated by using the trials and the address value of the cumulative matrix with respect to the output matrix obtained by calculating the repeat index. The new matrix is hence generated by first determining the largest of the both matrices and estimating the sum of the largest matrix and adding it with one. This is the new matrix obtained.

Perform mapping of the signal: The next step involes mapping of the obtained output matrix from the previous process. This process is performed to improve the effeciency of the speech signal.

Perform mutation: Mutation is the process of obtaining a new matrix by applying random changes to the older matrix.

Speech verification process: After performing the trial computation and mutation the difference between the test signal and each of the reference signals are computed again. Then the absolute value of the obtained output matrix is calculated and its sum is estimated. This is stored in a matrix. In the final step mapping is performed and the minimum value is considered (which will be zero).The minimum value and its location address are stored in two different arrays. Then the appropriate voice is

recognized and hence it gives the authentication. In this paper, speech samples of frequency less than 15000Hz are considered. Hence if the speech sample of frequency greater than 15000Hz are given, that signal is not authenticated.

III RESULTS

When the speech signal which is known as the test signal is given as the input, it is compared with that of the signals stored in the database which are known as reference or database signals. If the test signal (i.e., in the form of test matrix) matches with any of the reference signals (i.e., in the form of reference matrices), the speech signal is recognized and hence is authenticated.

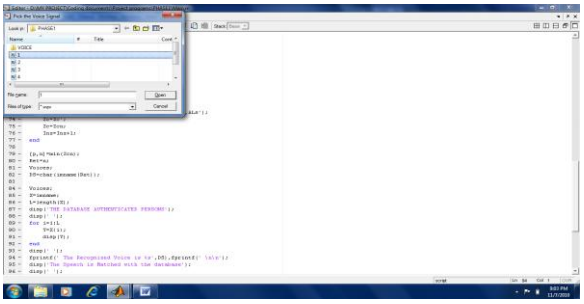

Fig 7 Giving the input signal

In the absence of noisy environments, the speech signal is given as the input. It is shown in Fig 7.

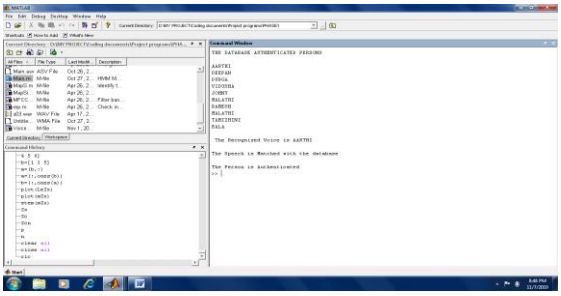

Fig 8 Displaying the speech authentication (positive case)

Figure 8 shows the result that the speech signal is matched with one of the signals stored in the database. In this project speech samples of 10 speakers are stored in the database. The names of those speakers are listed. Then the name of the recognized voice is described and it is matched with the database. Hence the speaker is authenticated.

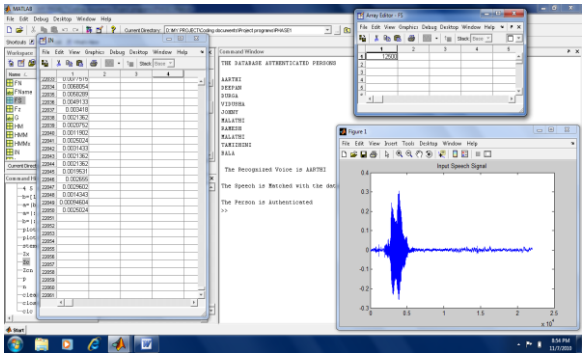

Figure 9 Representation of test signal ,test matrix and the sampling frequency of the test signal

Figure 9 represents the graphical plot of the test signal, and its matrix views. The test matrix is the representation of the input signal in matrix format. It is of the dimension 22050 x 1. The sampling frequency is 12,500 Hz. It is of dimension 1 x 1.

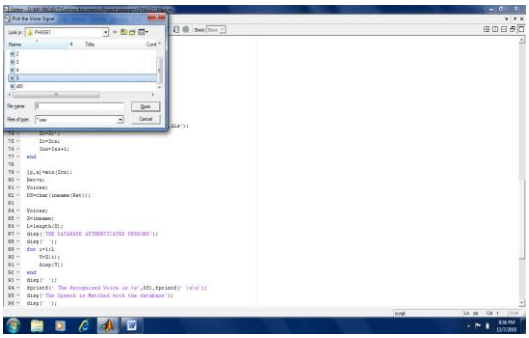

Fig 10 Giving the input signal for the negative case

Figure 10 shows the giving the speech signal for recognition. This is the case for the speaker not authenticated.

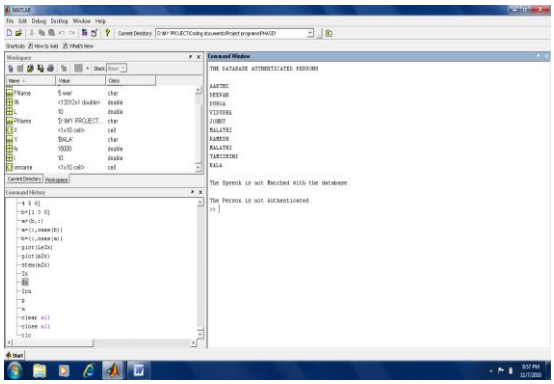

Figure 11 Displaying the speech authentication (negative case)

Figure 11 shows the result that the speech signal is not matched with those of the signals stored in the database. In this project speech samples of 10 speakers are stored in the database. The

name of those speakers are listed. Then the result is displayed that the input speech signal is not matched with the database and hence the speaker is not authenticated.

IV CONCLUSION

Hence, in the absence of noisy environments the features of n number of speech signals are extracted and stored it in the database, and the comparison is made with that of the input speech signal given at real time for recognition and thus to authenticate that speech signal. Speaker verification is one of the few recognition areas where machines can outperform humans. Speaker verification technology is a viable technique currently available for applications. Speaker verification can be augmented with other authentication techniques to add further security. This project is hence should be tested for speech recognition in the presence of noisy environments. The speech signal when given at real time is tested for de-noising. In the presence of noisy environments, de-noising is performed and the accuracy of speech recognition is to be viewed.

REFERENCES

[1]Mahadi Shaneh and Azizollah Taheri, "Voice Command Recognition System Based on MFCC and VQ Algorithms"

[2]Mahmoud I. Abdalla and Hanaa S.Ali, "Wavelet- Based Mel-Frequency Cepstral Coeffecients for Speaker Identification using Hidden Markov Models"

[3]Benjamin J. Shannon and Kuldip K. Paliwal, "MFCC Computation from Magnitude spectrum of Higher Lag Autocorrelation Co-effecients for Robust Speech Recognition"

[4]J.J. Gonzalez de la Rosa ,A.Moreno, A.Gallego,R.Piotrkowski, E.Castro, and J.Vico, "A Virtual Instrument for Acoustic Termite Detection Based in the Spectral Kurtosis"

[5] Jiehua Dai, Zhengzhe Wei, "Study and Implementation of Feature Extraction and Comparison in Voice Recognition"

[6] Yao Xie, Xiayu Zheng, "A Speaker Verification System"

[7]Bojan Kotnik, Damjan Vlaj, Zdravko Kacic Bogomir Horvat, "Robust MFCC Feature Extraction Algorithm using Efficient Additive and Convolutional Noise Reduction Procedures"