IMPROVEMENT OF THE TEXT DEPENDENT SPEAKER IDENTIFICATION SYSTEM USING DISCRETE MMM WITH CEPSTRAL BASED FEATURES

[1]Md. Rabiul Islam, [2]Md. Fayzur Rahman and [3]Muhammad Abdul Goffar Khan
[1]Department of Computer Science & Engineering
[2]Department of Electrical & Electronic Engineering
[3]Department of Electrical & Electronic Engineering
Rajshahi University of Engineering &
Technology (RUET), Rajshahi-6204, Bangladesh.
E-mail: [1]rabiul_cse@yahoo.com, [2]mfrahman3@yahoo.com, [3]qmagk@yahoo.com

Abstract: *In this paper, an improved strategy for automated text based speaker identification scheme has been proposed. The identification process incorporates the Hidden Markov Model technique. After preprocessing the speech, HMM is used in the learning and identification. Features are extracted by different techniques such as RCC, MFCC, ΔMFCC, ΔΔMFCC, LPC and LPCC which is almost different in each case. The highest identification rate of 93% has been achieved in the close set text dependent speaker identification system.*

Keywords: *Biometric Technologies, Automatic Speaker Identification, Cepstral Coefficients, Feature Extraction, Hidden Markov Model.*

1. Introduction

Biometrics is seen by many as a solution to a lot of user identification and security problems now-a- days [1]. Speaker identification is one of the most important areas where biometric techniques can be used. There are various techniques to resolve the automatic speaker identification problem. [2,3,4,5,6,7,8] A wide range of speaker identification applications are feasible over dialing-up telephones, including automation of operator assisted services, inbound and outbound telemarketing, call distribution by voice, expanded utility of a rotary phone, repertory dialing and catalog ordering. It is also used in voice controlled and operated games and toys, voice recognition aids for the handicapped and voice control of non strategic functions in a moving vehicle. The three important techniques for speaker identification are frequently used. They are the (i) acoustic-phonetic approach, (ii) the pattern recognition approach and (iii) the artificial intelligence approach. This paper deals with the pattern recognition approach.

In this work, close-set text dependent speaker identification technique has been considered and Discrete Hidden Markov Model has been used as a classification technique. The overall work has been simulated using MATLAB based toolbox such as Signal processing Toolbox, Voicebox and HMM Toolbox.

2. Paradigm of Speaker Identification System

The basic building blocks of speaker identification system are shown in the Figure1. The first step is the acquisition of speech from speakers. Then the start and end points of speech are detected. After which, pre-emphasis filtering technique has been used. The speech signal is segmented into overlapping analysis frames. After segmentation, windowing technique has been applied. Features are extracted from the segmented speech. The extracted features are then fed to the DHMM for learning and classification.

Implementation of the speaker identification system can be subdivided into two parts, (i) the speech signal processing and (ii) the Hidden Markov Model which is used for classification.

Date of submission : 27.07.2009 Date of acceptance : 22.12.2010

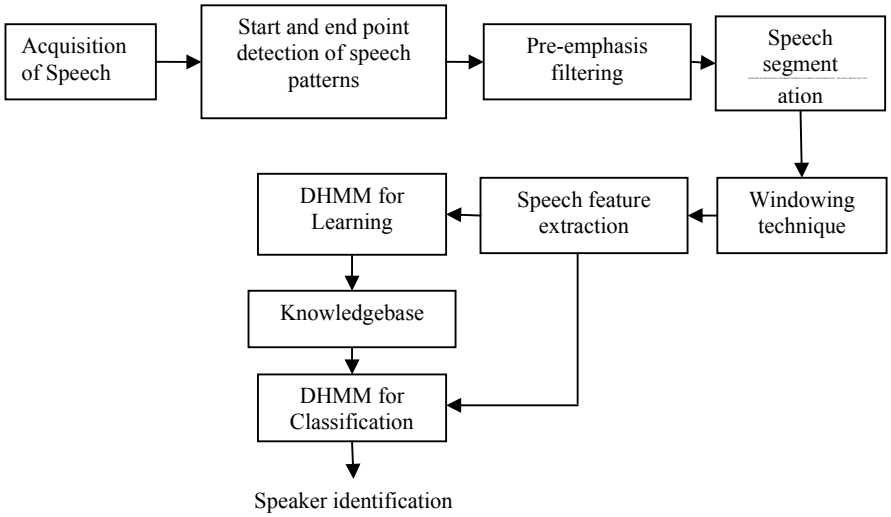


Fig. 1: Block Diagram of the proposed automated speaker identification system

3. Speech signal processing for speaker identification

3.1. Acquisition of Speech

Speech acquisition for this system has been done using high quality microphone in a sound proof room. To increase the accuracy of this system, it is necessary to keep speech acquisition process noise free. The speech data are recoded from 20 speakers. The length of the speech is about 3 seconds. Figure 2 shows a sample of the recorded speech. A sampling frequency of 11025 Hz with 16 bits resolution was used to record the speech voice. The recorded speech was saved in *.wav file format.

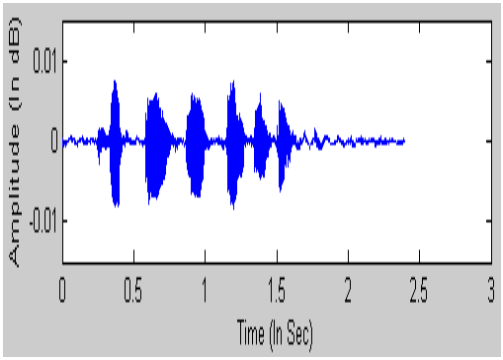


Fig. 2: Sample of the recorded speech signal

3.2 Start and End Point Detection

Speech end points detection algorithm has been used to detect the presence of speech, to remove pulse and silences in a background noise [9, 10, 11, 12]. Figure 3 shows the result after applying this algorithm over a speech signal.

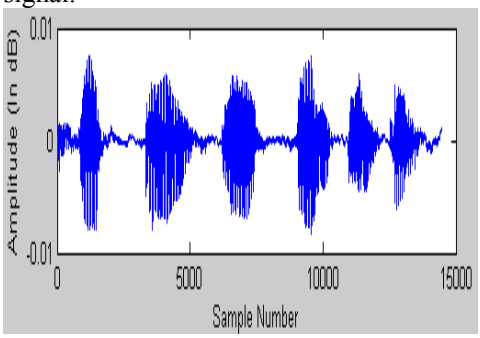


Fig. 3: Detection of the necessary speech information using start and end point detection algorithm

3.3 Pre-emphasizing

Pre-emphasis refers to filtering that emphasizes the higher frequencies. Pre-emphasis has been used to balance the spectrum of voiced sounds that have a steep roll-off in the high frequency region [13, 14, 15]. Figure 3 displays the output after applying the pre-emphasis filtering technique by using the equation:

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

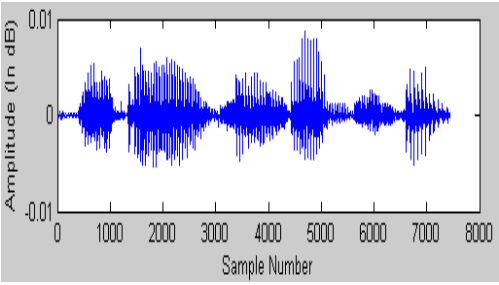


Fig. 4: Speech after Pre-emphasis filtering

3.4 Segmentation or Frame Blocking

In this step the continuous speech signal has been blocked into frames of N samples, with adjacent frames being separated by M (M < N). The first frame consists of the first N samples. The second frame begins M samples after the first frame and overlaps it by N-M samples. Similarly, the third frame begins 2M samples after the first frame (or M samples after the second frame) and overlaps it by N-2M samples. This process continues until all the speech is accounted. Typically a frame length of 10-30 milliseconds is used. A typical frame overlap is around 25% to 75% of the frame size. The purpose of the overlapping analysis is that each speech sound of the input sequence would be approximately centered at some frame [16, 17]. Figure 5 shows a segmented speech signal.

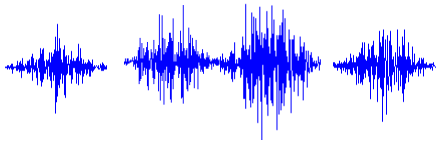


Fig. 5: Segmented speech

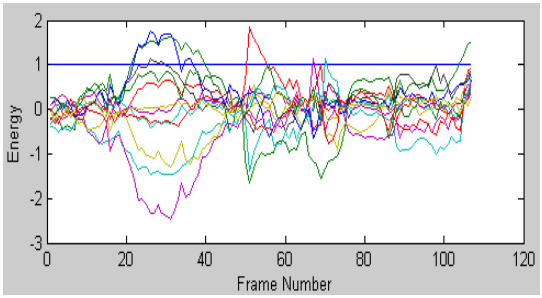
3.5 Windowing

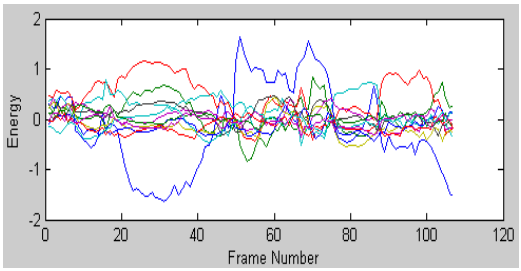
In this work, the purpose of using window is to reduce the effect of the spectral artifacts that results from the framing process [18, 19, 20]. From different types of windowing techniques, Hamming window has been chosen for this system. The hamming window has been implemented by the equation [18]:

$$w[k+1] = 0.54 - 0.46\cos\left(2\Pi\frac{k}{n-1}\right),\ k = 0,1,\ldots,n-1 \quad (2)$$

4. Featuer Extraction

This stage is very important in an ASIS because the quality of the speaker modeling and pattern matching strongly depends on the quality of the feature extraction methods. For the proposed ASIS, different types of speech feature extraction methods [21, 22, 23, 24, 25, 26] such as RCC, MFCC, ΔMFCC, ΔΔMFCC, LPC, LPCC have been applied. Figure 6 shows the features after applying different types of feature extraction techniques.

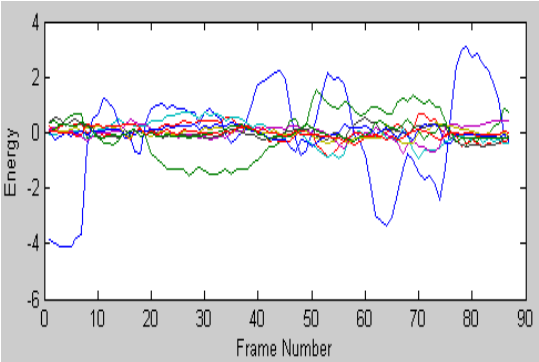


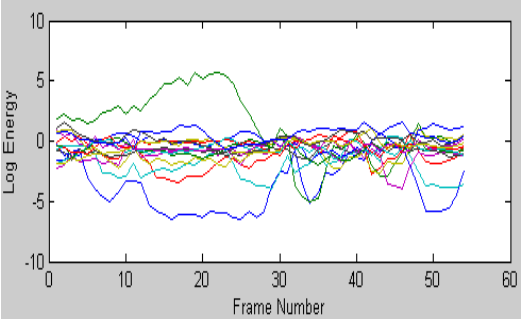


(a) LPC features of 12th order
(b) LPCC features of 12th order

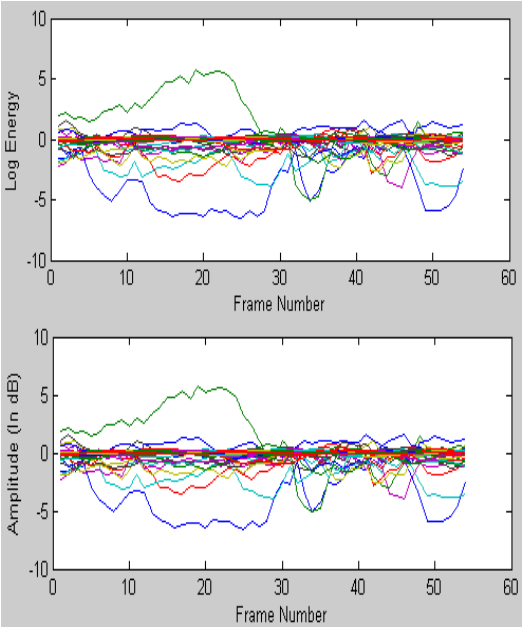
5. Feature Conditioning

Since DHMM can take only positive integer values as input, so it is required to transform the continuous valued features into discrete valued features. It can be performed by using vector quantization method. Vector quantization is a system for mapping a sequence of continuous or discrete vectors into a discrete codebook index. The results after applying feature conditioning are shown in Figure 7.





(c) RCC features with 15 coefficients
(d) MFCC features with 15 coefficients



(e) ΔMFCC features with 15 coefficients
(f) ΔΔMFCC features with 15 coefficients

Fig. 6: Feature extraction form of speech

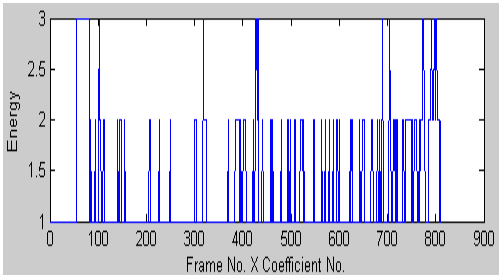


Fig. 7: Feature conditioning output

6. Speaker modeling

For each speaker k, an erogodic DHMM (Discrete HMM), θ_k has been built [27, 28, 29]. The model parameters (A, B, θ) have been estimated to optimize the likelihood of the training set observation vector for the k^{th} speaker by using Baum-Welch algorithm. The Baum-Welch re-estimation formula has been considered as follows [30]:

$$\bar{\Pi}_i = \gamma_1(i) \qquad (3)$$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \qquad (4)$$

$$\bar{b}_j(\vec{k}) = \frac{\sum_{t=1(s,t,\vec{o}_t=\vec{v}_k)}^{T} \gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \qquad (5)$$

where,

$$\xi_t(i,j) = \frac{\alpha_t(i) a_{ij} b_j(\bar{o}_{t+1}) \beta_{t+1}(j)}{\sum_{i=1}^{N}\sum_{j=1}^{N} \alpha_t(i) a_{ij} b_j(\bar{o}_{t+1}) \beta_{t+1}(j)} \text{ and}$$

$$\gamma_t(i) = \sum_{j=1}^{N} \xi_t(i,j)$$

In the testing phase, for each unknown speaker to be recognized, the processing shown in Figure 8 has been carried out. This procedure includes:

- Measurement of the observation sequence $O = \{o_1, o_2, ...o_n\}$, via a feature analysis of the speech corresponding to a speaker.

- Transformation of the continuous values of *O* into integer values.
- Calculation of model likelihoods for all possible models, $P(O \mid \theta_k), 1 \le k \le K$.
- Declaration of the speaker as *k**speaker whose model likelihood is highest, that is,

$$k^* = \arg\max_{1 \le k \le K} [P(O \mid \theta_k] \quad (6)$$

In this proposed work the probability computation step has been performed using the Baum's Forward-Backward algorithm [30, 31].

7. Experimental Result and Performance Analysis

There are some critical parameters (such as frame length, frame increment, number of cepstral coefficients, number of hidden states, pre-emphasizing parameter etc) that affect the performance of DHMM based close-set text-dependent speaker identification system. The optimal values of the above parameters are chosen to finalize the result.

7.1 Experiment on the window shift N_1

In this experiment, the effect of shifting of hamming window has been measured. By setting the window length, N_L = 15 ms, number of Mel-frequency Cepstral Coefficients excluding 0^{th} coefficients, N_{MC} =15, number of hidden states, N_H =5 and the emphasizing parameter, α = 0.9, we have found the highest speaker identification rate of 85[%] is at 75% window shift as shown in Figure 9.

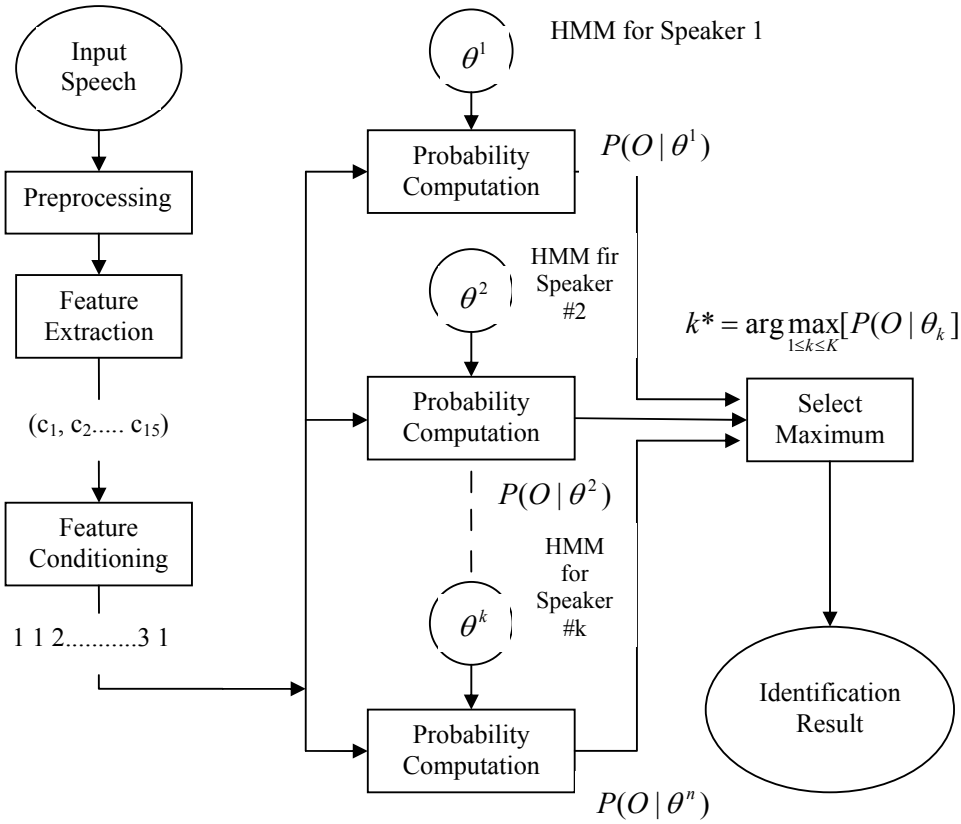


Fig. 8: Block diagram of speaker DHMM recognizer

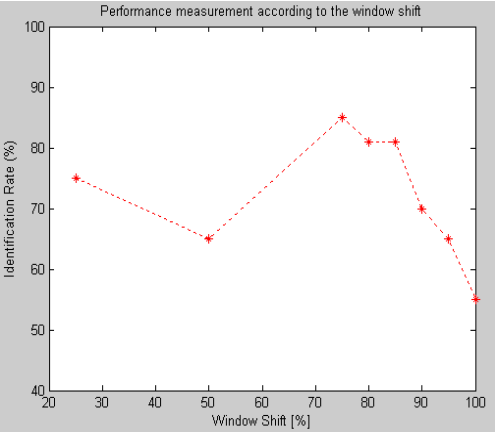


Fig. 9: Performance measurement according to the window shift

7.2 Experiment on the Pre-emphasized parameter, α

In this experiment, the performance of the developed speaker identification system has been measured according to the pre-emphasized parameter α. We have set N_L = 15 ms, N_I =15 ms, N_{MC} =15 and N_H =5. We have studied the value of the parameter ranges from 0.7 to 0.99. We have found that the speaker identification performance was 85[%] at α = 0.95 which is shown in Figure 10.

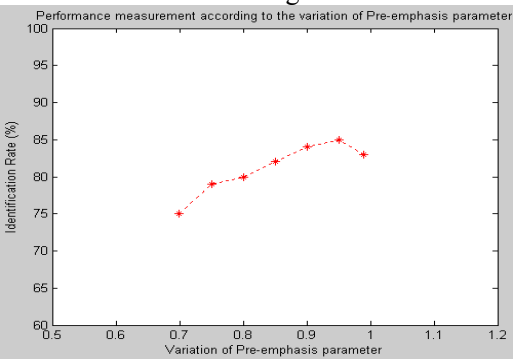


Fig. 10: Speaker identification rate on the variation of pre-emphasis parameter

7.3 Experiment of the number of hidden states of DHMM, N_H

In the learning phase of DHMM, The hidden states have been chosen in the range from 5 to 20. We have set N_L = 15 ms, N_I =15 ms, N_{MC} =15, and α = 0.95. The highest performance has been achieved at N_H =15 which is shown in Figure 11.

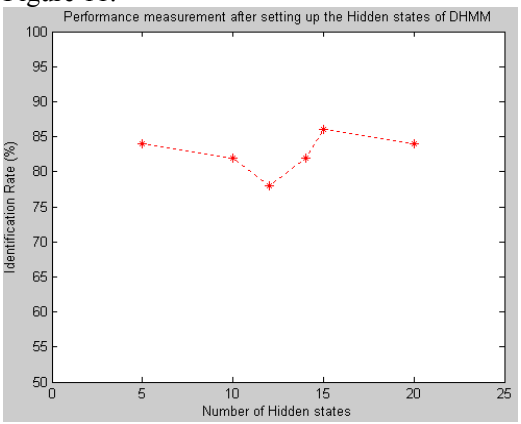


Fig. 11: Results after setting up the hidden states of DHMM

7.4 Effects of the window length, N_L

We have chosen the window length, N_L from 10 ms to 30 ms. By setting N_L = 15 ms, N_I =15 ms, N_{MC} =15 and α = 0.95, the highest performance of 87[%] has been achieved at MFCC based system. Figure 12 shows the result.

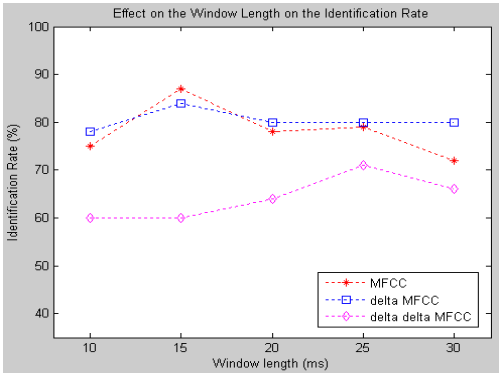


Fig. 12: Effect of the window length on the identification rate

7.5 Effects of the number of cepstral coefficients, N_C

In this experiment, the number of cepstral coefficients was varied from 10 to 20 with a step size 2. According to the parameters at N_L = 15 ms, N_I =15 ms, N_{MC} =15 and α = 0.95, the highest speaker identification rate was 93[%] which was achieved for ΔMFCC per frame.

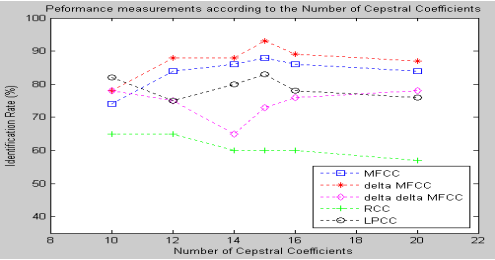


Fig. 13: Speaker identification accuracy according to the number of cepstral coefficients

8. Conclusion and Observations

The critical parameters such as frame length, frame increment, number of cepstral coefficients, number of hidden states and the emphasizing parameter have a great impact of the identification performance of a DHMM based close set text dependent ASIS. To find out the best performance of this system, the optimal values of the above parameters have been selected effectively. Five experiments have been performed for this purpose. The highest identification rate of 93[%] has been achieved at ΔMFCC. Since the highest speaker identification rate was 93[%], this can satisfy the practical demand. The performance of this system can also be improved by the improvement of speech signal processing part and by using the hybrid system. Open set text independent speaker identification system with noisy speech can be the further work of this system.

References

[1] A. Jain, R. Bole, S. Pankanti "BIOMETRICS Personal Identification in Networked Society" Kluwer Academic Press, Boston, 1999.

[2] Rabiner, L., and Juang, B.-H., Fundamentals of Speech Recognition, Prentice Hall, Englewood Cliffs, New Jersey, 1993.

[3] Jacobsen, J. D., "Probabilistic Speech Detection," Informatics and Mathematical Modeling, DTU, 2003.

[4] Jain, A., R.P.W.Duin, and J.Mao., "Statistical pattern recognition: a review", IEEE Trans. on Pattern Analysis and Machine Intelligence 22 (2000), 4–37.

[5] Davis, S., and Mermelstein, P., "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," IEEE 74 Transactions on Acoustics, Speech, and Signal Processing (ICASSP), vol. 28, no. 4, pp. 357-366, Aug. 1980.

[6] Sadaoki Furui, "50 Years of Progress in Speech and Speaker Recognition Research", ECTI TRANSACTIONS ON COMPUTER AND INFORMATION TECHNOLOGY Vol.1, No.2, November 2005.

[7] Lockwood, P., Boudy, J., and Blanchet, M., "Non-linear spectral subtraction (NSS) and hidden Markov models for robust speech recognition in car noise environments," IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), vol. 1, pp. 265-268, Mar. 1992.

[8] Matsui, T., and Furui, S., "Comparison of text-independent speaker recognition methods using VQ-distortion and discrete/ continuous HMMs," IEEE Transactions on Speech Audio Process, no. 2, pp. 456-459, 1994.

[9] Koji Kitayama, Masataka Goto, Katunobu Itou and Tetsunori Kobayashi, "Speech Starter: Noise-Robust Endpoint Detection by Using Filled Pauses", Eurospeech 2003, Geneva, pp. 1237-1240.

[10] S. E. Bou-Ghazale and K. Assaleh, "A robust endpoint detection of speech for noisy environments with application to automatic speech recognition", in Proc. ICASSP2002, vol. 4, 2002, pp. 3808–3811.

[11] A. Martin, D. Charlet, and L. Mauuary, "Robust speech / non-speech detection using LDA applied to MFCC", in Proc. ICASSP2001}, vol. 1, 2001, pp. 237–240.

[12] Richard. O. Duda, Peter E. Hart, David G. Strok, "Pattern Classification", A Wiley-interscience publication, John Wiley & Sons, Inc, Second Edition, 2001.

[13] Harrington, J., and Cassidy, S. Techniques in Speech Acoustics. Kluwer Academic Publishers, Dordrecht, 1999.

[14] Makhoul, J. Linear prediction: a tutorial review. Proceedings of the IEEE 64, 4 (1975), 561–580.

[15] Picone, J. Signal modeling techniques in speech recognition. Proceedings of the IEEE 81, 9 (1993), 1215–1247.

[16] Clsudio Beccchetti and Lucio Prina Ricotti, Speech Recognition Theory and C++ Implementation, John Wiley & Sons. Ltd., pp.124-136.

[17] L.P. Cordella, P.Foggia, C. Sansone, and M. Vento, "A Real-Time Text-Independent Speaker Identification System", Proceeding of the 12th Intern ational Conference on

[18] J. R. Deller, J. G. Proakis, and J. H. L. Hansen. Discrete-Time Processing of Speech Signals. Macmillan, 1993.

[19] F. Owens. Signal Processing Of Speech. Macmillan New electronics. Macmillan, 1993.

[20] F. Harris, "On the use of windows for harmonic analysis with the discrete fourier transform", Proceedings of the IEEE 66, vol.1 (1978), pp.51-84

[21] D. Kewley-Port and Y. Zheng. Auditory models of formant frequency discrimination for isolated vowels. Journal of the Acostical Society of America, 103(3):1654–1666, 1998.

[22] D. O'Shaughnessy. Speech Communication - Human and Machine. Addison Wesley, 1987.

[23] E. Zwicker. Subdivision of the audible frequency band into critical bands (frequenzgruppen). Journal of the Acoustical Society of America, 33:248–260, 1961.

[24] S. Davis and P. Mermelstein. Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. IEEE Transactions on Acoustics Speech and Signal Processing, 28:357–366, Aug 1980.

[25] S. Furui. Speaker independent isolated word recognition using dynamic features of the speech spectrum. IEEE Transactions on Acoustics, Speech and Signal Processing, 34:52–59, Feb 1986.

[26] S. Furui, "Speaker-Dependent-Feature Extraction, Recognition and Processing Techniques." Speech Communication,Vol. 10, pp. 505-520, 1991.

[27] Huang, X. D., Ariki, Y., and Jack, M. A., 1990. Hidden Markov Models for Speech Recognition. Edinburgh University Press, Scotland, UK.

[28] Hwang**,** M**.,** and Huang**,** X., 1993. Shared**-**Distribution Hidden**.** Markov Models for Speech Recognition. IEEE. Trans. on. Speech and Audio Processing, vol. 1, No. 4, pp. 414-420.

[29] Baum**,** L.E., Petrie, T**.,** Soules, G., and Weiss, N., 1970. A maximization technique occurring in the statistical analysis of probabilistic functions of Markov chains. The Annals of Mathematical Statistics, 41, pp. 164-171.

[30] Rabiner, L. R., 1989. A tutorial on hidden Markov models and selected applications in speech recognition. Proceedings of the IEEE, vol. 77, no. 2, pp. 257–286.

[31] Devijver, P. A., 1985. Baum's forward**-**backward algorithm revisited. Pattern Recognition Letter, 3, pp. 369-373.