International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(3), pp.303-310
DOI: http://dx.doi.org/10.21172/1.83.045
e-ISSN:2278-621X

CONNECTED SPEECH RECOGNITION FOR AUTHENTICATION

Sheena Christabel Pravin[1] and S.Satheesh Kumar[2]

Abstract- **In today's fast world we don't have time to sit and type our complicated alpha-digit codes used for Product key, for Password Security Code (PSC) to access a person's authenticated details containing unique set of connected alpha-digit. Recognition of spoken alphabets and digits is difficult task in automatic speech recognition due to phonetic similarities among certain group of vocabulary sets. In this paper, knowledge based features are used along with conventional Mel Frequency Cepstral Coefficients (MFCC) to enhance the overall correctness of the PSC recognizer by overcoming phonetic similarity difficulties. HMM (Hidden Markov Model) is used as the statistical classifier which determines the likelihood of every speech sample. In this paper, we achieve speech to text conversion using MATLAB. It extracts and labels the waveform and gives output in the text format. Word Boundary detection followed by HMM based classification results in a PSC recognition accuracy of 83%.**

Keywords – Phonetics, MFCC, HMM, MATLAB, Word Boundary Detection

I. INTRODUCTION

Speech is the fundamental and effortless mode of communication. The process of converting speech data recorded by a microphone or a telephone into text is called speech recognition. Automatic Speech Recognition (ASR) technology permits a system to recognize the words that a person speaks into a microphone or telephone. Automatic recognition of spoken alphabet and digit is one of the difficult tasks in speech recognition. Spoken alpha-digit recognition finds its applications in spelled name and addresses recognition, telephone number recognition, product key, registered code or reference ID, Automation of directory assistance, credit card entry, Personal Identification Number (PIN) entry, entry of access codes for transactions and voice dialing etc. Spoken alphabet and digit recognition seems to be a simple task for human being. But for machine, this can be a challenging task due to high acoustic similarities among certain groups of letters [1]. In this paper, the development of an alphabet and digit recognition system which can be used for specific application is presented. The application considered here is recognition of Password Security Code (PSC).

PSC acts as a unique set of alpha-digit for a person maintaining his authenticated files in an organization. Automatic PSC recognition is useful in opening up a person's pdf documents which is password protected where, some other person can access his documents remotely with his permission but without knowing his password. The PSC follows a specific format and it consists of both alphabets and digits. Therefore, recognition of PSC is considered to be an application of connected alpha-digit recognition system. While designing alpha-digit speech recognizer, the phonetic similarities among E-set alpha-digit vocabulary (B,C,D,E,G,P,T,V,Z,3), the A-set (A,H,J,K,8) and the nasal set (M,N) presents difficulties in recognition [2]. Sufficient works has been made in the area of alpha-digit recognition in different languages and different degrees of accuracies were obtained with different algorithms. In literature, Mashao et al. [2] made a comparative study for alpha-digit recognition using both Linear Predictive Co-efficient

[1] *Department of Electronics and Communication Engineering Rajalakshmi Engineering College, Chennai, Tamil Nadu, India.*
[2] *Department of Electronics and Communication Engineering Rajalakshmi Engineering College, Chennai, Tamil Nadu, India.*

(LPC) and Discrete Fourier Transform (DFT) features and found that DFT feature out performs over LPC. Hamaker et al. [3] proposed an alpha-digit recognition system using context-independent syllables, and reported that performance of syllable based system, which simultaneously exploits temporal and spectral variations, was better than phone based system for Large Vocabulary Continuous Speech Recognition (LVCSR). Hemalatha et al. [4] used an unsupervised and incremental training procedure, which does not require segmented and labeled speech corpora for connected digit recognition. Lee [5] modeled duration and spectral dynamics of speech signal as duration high-order Hidden Markov Model (HMM), to improve the continuous Mandarin digit recognition. Grag et al. [6] used MFCC and Dynamic Time Warping (DTW) for connected digit recognition. Martino [7] used feature vector containing up to fifth order derivatives to enhance the alphabet recognition performance. Adam and Salam [8] used MFCC speech feature and Feed Forward Back Propagation Neural Network (FFBPNN) with adaptive learning rate to classify the highly confusable E-set alphabets those share the same /iy/ vowel at the back end of its utterance. In this paper, [9] the Hidden Markov Model Toolkit (HTK) is used for developing the PSC recognition system.

The baseline system is initially designed as a phoneme level recognizer. The silence (sil) model is also included in the model set. Since, most of the digits consisted of more than two phonemes, context-dependent tri phone models are created from the mono phone models. Initially, MFCC speech features are used. Later, other knowledge based features are added along with MFCC to overcome confusions due to phonetic similarities among particular vocabulary set and to enhance the recognizer performance. [10] Similar to MFCC we use word detection boundary in order to detect both additive noise and noise-induced changes in the talker's speech production. In the next Section, the baseline USN recognizer is presented. Additional features used along with MFCC are described in Section III. Results are included in Section IV, followed by conclusion of the work in Section V.

II. FEATURE EXTRACTION

Mel Frequency Cepstral Co-efficients (MFCC) closely approximates the human auditory system's response. It is based on a linear cosine transform of a log power spectrum on a nonlinear Mel-scale in frequency domain. Mel scale is linear up to 1 kHz frequency and then behaves logarithmically. In advancement to the Fast Fourier Transform(FFT) parameters which were earlier used for feature extraction, in MFCC, the frequency bands are positioned logarithmically (on Mel scale) rather than being linearly spaced in case of FFT. The flow diagram of MFCC is described in Figure 2

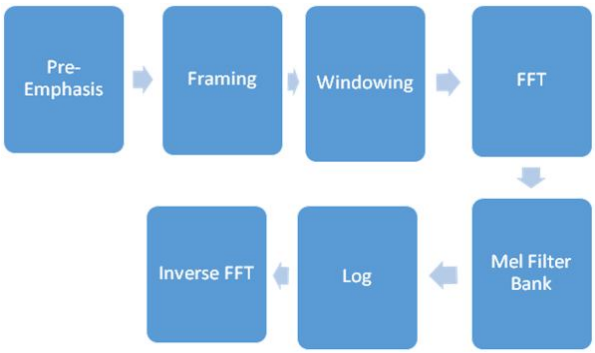

Figure 1. MFCC Feature Extraction

III. WORD BOUNDARY DETECTION

When trying to incorporate an isolated word recognizer to recognize connected speech, the most crucial part is the separation of the incoming signal into the constituent words. This is extremely difficult in noisy environments. It is particularly troublesome with words that begin or end in low-energy phonemes or with words which have a silence before release [11]. Some speakers also habitually allow their words to trail off in energy. Others tend to produce bursts of breath noise at the end of the words [3]. In the conventional endpoint detection algorithms, the short-time energy or spectral energy is usually used as the primary feature parameters with the augmentation of zero-crossing

rate, pitch and duration information. But these features become less reliable in the presence of non-stationary noise and various types of sound artifacts [12-14].

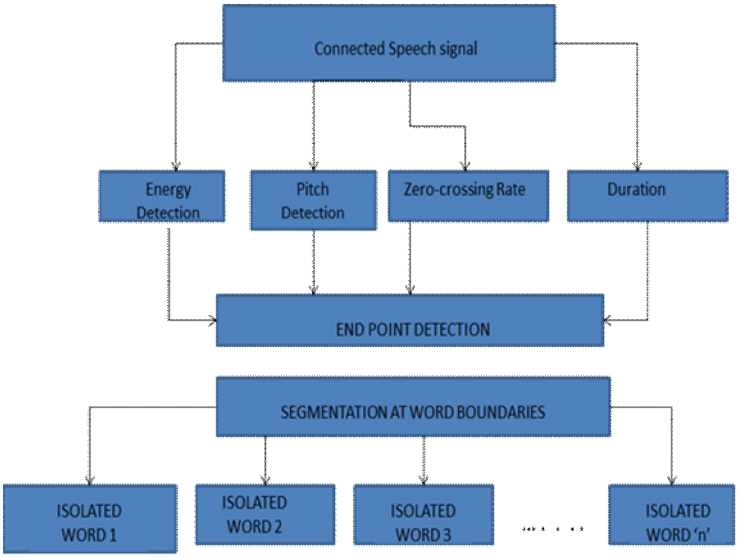

Figure 2: Proposed Model for Connected Speech Recognition

IV. HIDDEN MARKOV MODEL FOR PATTERN CLASSIFICATION

A. Elements of HMM

We now formally define the elements of an HMM, and explain how the model generates observation sequences. An HMM is characterized by N, the number of states in the model. Although the states are hidden, for many practical applications there is often some physical significance attached to the states or to sets of states of the model. Generally, the states are interconnected in such a way that any state can be re ached from any other state (e.g., an ergodic model). We denote the individual states as S = {Sl, S2, . . .,SN}, and the state at time t as qt. 'M' is the number of distinct observation symbols per state, i.e., the discrete alphabet size. The observation symbols correspond to the physical output of the system being modeled. We denote the individual symbols as

$$V = \{V1, V2\ldots\ldots\ldots, VM\} \quad (1)$$

The state transition probability distribution A = $\{a_{ij}\}$ where for the special case where any state can reach any other state in a single step, we have $a_{ij}>0$ for all i, j. For other types of HMMs, we would have aij= 0 for one or more (i, j) pairs. The observation symbol probability distribution in state j, B = {bj(k)}, where

$$Bj(k) = p\,[vk \text{ at } t|qt = Sj],\ 1 \le j \le N,\ 1 \le k \le M \quad (2)$$

A Markov chain with 5 states with selected state transitions is described in the Figure 3. A full probabilistic description of this system would, in general, require specification of the current state (at time t), as well as all the predecessor states. For the special case of a discrete, first order, Markov chain, this probabilistic description is truncated to just the current and the predecessor state, i.e.,

$$P\,[\,qt] = Sj|qt\text{-}1 = Si,\ qt\text{-}2 = Sk\ldots] \quad (3)$$
$$= P[qt = Sj|qt\text{-}1 = Si]$$

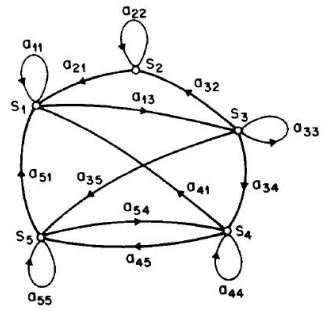

Figure 3: A Markov chain with 5 states with selected state transitions

Furthermore, we only consider those processes in which the right-hand side of (1) is independent of time, thereby leading to the set of state transition probabilities a_{ij} of the form

$$a_{ij} = P[qt] = {}_{Sj}\,[qt\text{-}1 = Si].\ 1 \leq i, j \leq N \qquad (4)$$

The above stochastic process could be called an observable Markov model since the output of the process is the set of states at each instant of time, where each state corresponds to a physical (observable) event.

B. Isolated Word Recognition

Assume we have a vocabulary of V words to be recognized and that each word is to be modeled by a distinct HMM. Further assume that for each word in the vocabulary we have a training set of K occurrences of each spoken word (spoken by 1 or more talkers) where each occurrence of the word constitutes an observation sequence, where the observations are some appropriate representation of the (spectral and/or temporal) characteristics of the word. For each word v in the vocabulary, we must build HMM λ_v, i.e., we must estimate the model parameters (A, B, Π) that optimize the likelihood of the set observation vectors for the vth word. For each unknown word which is to be recognized, the processing shown below must be carried out, namely measurement of the observation sequence ,0 = {O1, O2……OT}, via a feature analysis of the speech corresponding to the word; followed by calculation of model likelihoods for all possible models, $P(O|\lambda_v)$, $1 \leq v \leq V$; followed by selection of the word whose model likelihood is highest, i.e.,

$$V^* = \frac{\arg\max}{1 \leq v \leq V} [P(O/\lambda^V]) \qquad (5)$$

The probability computation step is generally performed using the Viterbi algorithm. To find the single best state sequence, Q = {ql q2, . . . qr}, for the given observation sequence O ={O1,O2,O3… OT) we need to define the quantity

$$\delta_t^i = \frac{\max}{(q1,q2,\ldots\ldots qt-1)} P[q1\ q2\ \ldots\ldots qt = i, \frac{O_1 O_2 \ldots\ldots O_t}{\lambda}] \qquad (6)$$

i.e., δ_t^i is the best score (highest probability) along a single path, at time t, which accounts for the first t observations and ends in state Si. By induction we have

$$\delta_{t+1}(j) = [\max_i \delta_t(i)\, a_{ij}].b_j(O_{t+1}) \qquad (7)$$

To actually retrieve the state sequence, we need to keep track of the argument which maximized the above equation for each t and j. We do this via the array ψt(j).

IV. RESULTS AND DISCUSSIONS

The material presented in this paper describes the implementation of a speaker dependent speech recognition system to recognize strings of words using continuous Hidden Markov Model (HMM). For now, the vocabulary consists of the ten digits (0-9) and alphabets (A-Z) of the English language. The main motivation for this is to incorporate a voice recognition system into a hands-free telephone system. The speech recognition system is implemented on MATLAB. Connected-word recognition system provides a more comfortable interface when implemented in computer interaction systems since they allow the users to speak in complete sentences. Zero Crossing Rate

indicates the presence of appreciable signal in the region of analysis. Word Boundaries need to be identified for individual recognition of the connected words. Figure 4 depicts the endpoints in the connected test speech signal.
The short-term energy function as shown in Figure 5 is a positive function and can therefore be processed in a manner similar to that of the magnitude spectrum to be used for automatic word boundary detection and segmentation. For segmenting the speech signal at syllable boundaries, energy based Word Boundary detection methods are good. As the segment structure is preserved in the short-term energy, in spite of noise. Since the alphabets and digits used in this project are one-syllable words, short term energy is an effective parameter for Word Boundary detection.

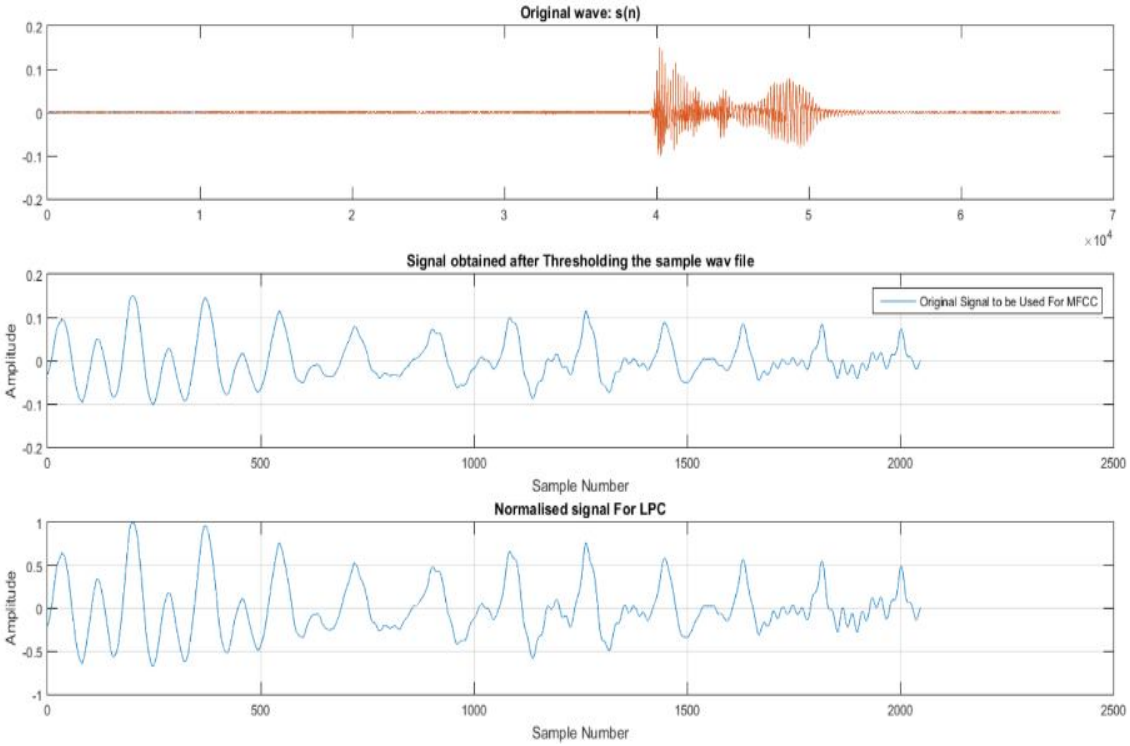

Figure 4: Zero Crossing Rate

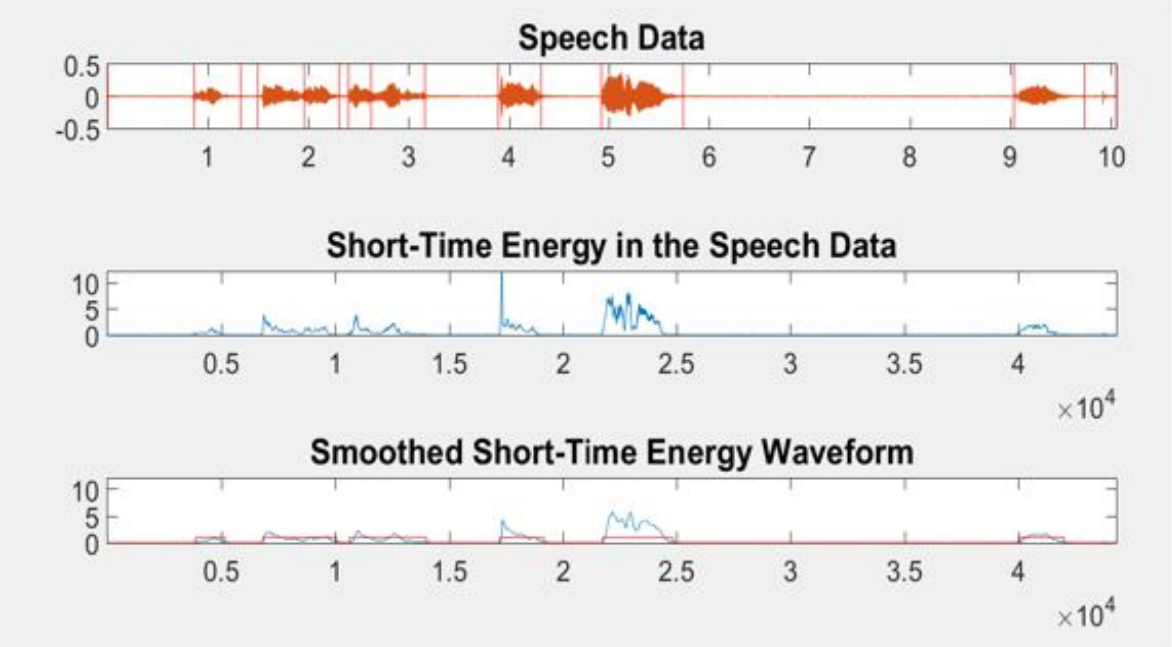

Figure 5: Automatic Word Boundary Detection

The HMM is the most popular and successful stochastic approach to speech recognition in general use. In the training phase, state transition probability distribution, observation symbol probability distribution and initial state probabilities are estimated for each speaker as a speaker model. The probability of observations for a given speaker model is calculated for speaker recognition. HMMs are used in speech recognition because a speech signal can be viewed as a piecewise stationary signal or a short-time stationary signal. In a short time-scales (e.g., 10 milliseconds), speech can be approximated as a stationary process. Speech can be thought of as a Markov model for many stochastic purposes. HMMs are popular is because they can be trained automatically and are simple and computationally feasible to use. After training unique HMM model for each alphabet and digit, the following results have been obtained:

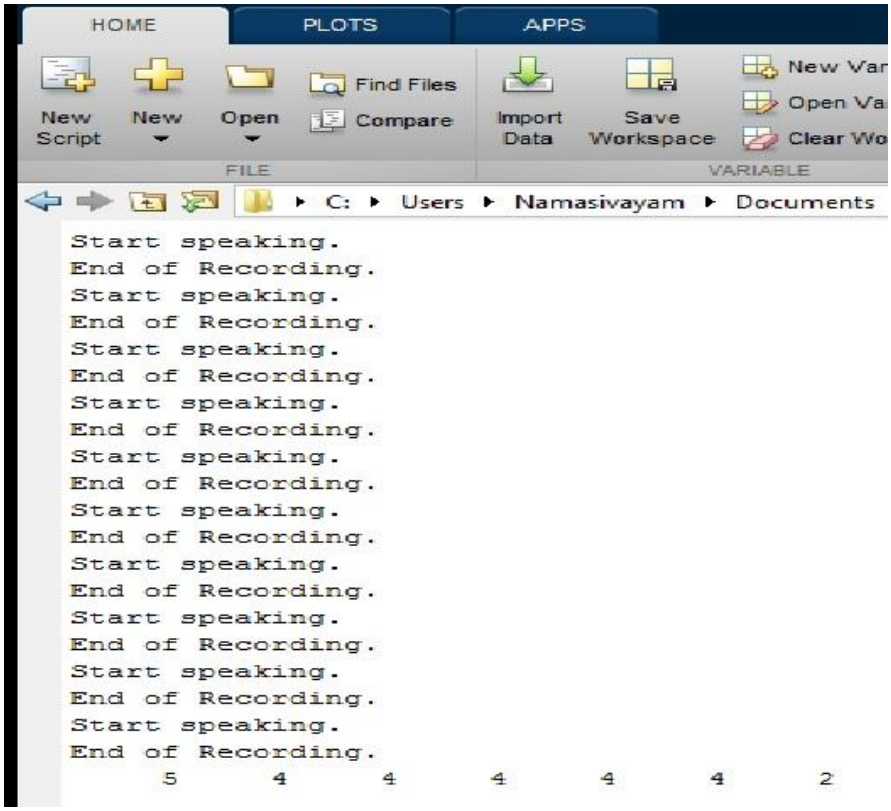

Figure 6: Automatic Speech Recognition Output

v. CONCLUSION AND FUTURE WORK

A connected speech recognizer using isolated word recognizer has been implemented for speaker dependent data. The efficiency of system to divide speech into words is showing appreciable results. Further the work can extended for speaker independent data by improving the efficiency of isolated word recognizer. The system is made robust to work in noisy environment. The real time system can be made by extending the database.

REFERENCES

[1] L.R. Rabiner, J.G. Wilpon, F. K. Soong, "High performance connected digit recognition", IEEE Trans. On Acoustics, Speech and Signal Processing, Vol.37, No.8, 1989.

[2] D.J. Mashao, Y. Gotoh and H. F. Silverman, "Analysis of LPC/DFT feature for an HMM-based alpha-digit recognizer", IEEE signal processing letters, Vol.3, No.4, pp. 103-106, 1996.

[3] J. Hamaker, A. Ganapathiraju, J. Picone, J.J. Godfrey, "Advances in alphadigit recognition using syllables", In Proc. IEEE International Conference on Acoustic, Speech and Signal processing (ICASSP), Vol.1, USA, 1998.

[4] N.Hemalatha, G.L. Sarada, T. Nagarajan, H.A Murthy, "Connected digit recognition using unsupervised and incremental HMMs", In Proc. of the Eleventh National Conference on Communication (NCC), Kharagpur, 2005.

[5] L.M. Lee, "High-Order Hidden Markov Model and Application to Continuous Mandarin Digit Recognition", Journal of Information Science and Engineering, Vol.27, pp.1919-1930, 2011.

[6] Amit Grag, Nikita and Poonam, "Connected Digits Recognition using Distance Calculation at each Digit", International Journal of Computational Engineering and Management, Vol. 14, No.1, pp. 94-96, 2011.

[7] J. Martino, "On the use of high order derivatives for high performance alphabet recognition", In Proc. IEEE International Conference on Acoustics, Speech, and Signal Processing, Florida, USA, 2002.

[8] T.B. Adam, M.D Salam, "Spoken English alphabet recognition with Mel Frequency Cepstral Coefficients and Back Propagation Neural Networks", International Journal of Computer Applications, Vol. 42, No.12, pp. 572-587, 2012.

[9] L. Rabiner "Introduction to Hidden Markov model", IEEE, pp.4-16, 2003.

[10] J.-C. Junqua, "A Robust algorithm for Word Boundary Detection in the presence of noise" , IEEE signal processing, pp.406-412, 2002.

[11] R.Deller Jr., John H.L. Hansen, John G. Proakis, "Discrete timeprocessing of speech signalsJohn", IEEE Press Classic Reissue, IEEE press, 2000.

[12] WANG Fan, ZHENG Fang, WU Wenhu “A self-adapting endpoint detection algorithm for speech recognition in noisy environments based on 1/f process”, Center of Speech Technology, State Key Laboratory of Intelligent Technology and Systems, Department of Computer Science & Technology, Tsinghua University, Beijing, 2000.

[13] Philippe Renevey and Andrzej Drygajlo, “Entropy Based Voice Activity Detection in Very Noisy Conditions”, Swiss Center for Electronics and Microtechnology, Neuchatel, Switzerland Swiss Federal Institute of Technology, Lausanne, Switzerland

[14] Jia-lin Shen, Jeih-weih Hung, Lin-shan Lee “Robust Entropy-based Endpoint Detection for Speech Recognition in Noisy Environments”, Institute of Information Science, Academia Sinica Taipei, Taiwan, Republic of China

[15]. Rabiner, L.R. and Juang, B.H. (1986), "An Introduction to Hidden Markov Models," iEEE ASSP MagAzine, Vol. 3, No. 1, pp. 4-17, Jan. 1986.