Implementation of HMM and Radial Basis Function for Speech Recognition

[1]Umarani S D,
Research Scholar(Full Time) ,
Dept. of Electronics and Communication Engg.,
Government College of Engg.,
Salem – 636011, Tamilnadu, INDIA
Email id : umaraviram@gmail.com
Mobile : 00-91-9843688818.

[2]Raviram P
Research Scholar,
Dept. of Computer Science Engg.,
Vinayaka Missions University
Salem, Tamilnadu – 636 308, INDIA
E:mail: ravirampedu@gmail.com
Tel: 00-91-9843968884

[3]Wahidabanu R S D
Professor and Head,
Dept. of Electronics and Communication Engg.,
Government College of Engg., Salem - 636 011, INDIA
Tel: 00-91-9443008886. E:mail: rsdwb@yahoo.com

***Abstract*—The work aims at recognizing words from a continuous speech. To achieve this, cepstrum analysis of the speech signal is carried out. The speech signal is processed and the features are extracted using cepstrum analysis. The extracted features are given as inputs for the hidden Markov model (HMM) followed by training radial basis function (RBF). During the testing process, the words are separated and compared in the database. If a word matches then subsequent action is carried out. If the word is not present, then it is added to the database.**

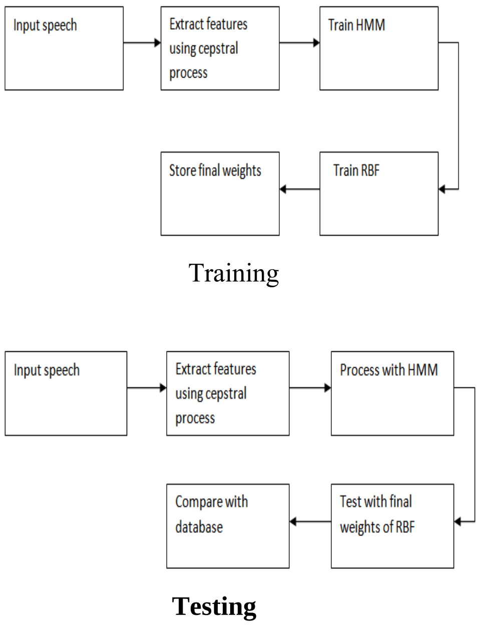
Keywords- Hidden Markov model; Radial basis function; cepstrum analysis; artificial neural network; speech recognition

I. INTRODUCTION

Automatic speech recognizing capability reduces drastically when noise is present [1]. To mitigate the effect of noise on recognition, noisy speech is typically preprocessed by speech enhancement algorithms, such as spectral subtraction based systems [2]. If samples of the corrupting noise source are available a priori, a model for the noise can additionally be trained and noisy speech may be jointly decoded based on the models of speech and noise [3]. However, in many realistic applications, the performance of the above approaches to robust speech recognition is inadequate. A missing data approach to robust speech recognition has been proposed [4]. This method distinguishes between reliable and unreliable data in the spectral or time-frequency (T-F) domain. When speech is contaminated by additive noise, some T-F regions will contain predominantly speech energy (reliable) and the rest are dominated by noise energy. The missing data method treats the latter T-F units as missing or unreliable during recognition. The performance of the missing data recognizer is significantly better than the performance of a system using spectral subtraction for speech enhancement followed by recognition of enhanced speech.

The missing data recognizer requires a binary T-F mask that provides information about which T-F regions, of the noisy speech signal, are reliable and which are unreliable. Previous studies have shown that the missing data recognizer performs exceedingly well when this mask is known a priori [5]. Attempts to estimate such a binary mask through front-end preprocessing using speech separation techniques have been only partly successful. Spectral subtraction is frequently used to generate such binary masks in missing data studies [6]. Noise is assumed to be long-term stationary and its spectrum estimated from frames that do not contain speech (silent frames containing background noise). The noise spectrum is then used to estimate the signal to noise ratio (SNR) in each T-F unit. If the SNR in a T-F unit exceeds a threshold, it is labeled reliable; it is labeled unreliable otherwise. In the presence of non-stationary interference sources, however, the use of spectral subtraction results in a poor estimate of the mask. Methods that primarily utilize the harmonicity of voiced speech have also been proposed to estimate the mask for missing data applications [7, 8, 9]. Hence, they are unable to effectively deal with unvoiced speech. Additionally, accurate estimation of pitch is difficult, if not impossible, when SNR is low. Under these conditions, estimation of the binary mask corresponding to voiced speech may not be reliable too. Thus, the estimation of the binary T-F mask remains a challenging problem.

978-1-4244-4711-4/09/$25.00 ©2009 IEEE


Input speech
Extract features using cepstral process
Train HMM
Train RBF
Store final weights
Input speech
Extract features using cepstral process
Process with HMM
Test with final weights of RBF
Compare with database

Training

Testing

Figure 1. SCHEMATIC DIAGRAM

The human auditory system exhibits a remarkable ability to segregate a target speech source from various interference. According to Bregman [10], this is accomplished via a process termed auditory scene analysis (ASA). ASA involves two types of organization, primitive and schema-driven. Primitive ASA is based on bottom-up cues such as pitch, and spatial location of a sound source. Schema-based ASA is based on top-down use of stored knowledge about auditory inputs, e.g. speech patterns, to supplement primitive analysis. In this paper, a cepstrum analysis method is proposed for feature extraction of the speech signal.

II. CEPSTRUM ANALYSIS

It is convenient to assume that the signal consists of a discrete time sequence, so that the spectrum consists of a z transform evaluated on the unit circle. Let us consider a speech example, with X referring to the spectrum of the observed speech signal, E to the excitation component (for instance, the glottal pulse train), and V to the vocal tract shaping of the excitation spectrum. We begin with a multiplicative model of the two spectra (the excitation and the vocal tract). Thus, the spectral magnitude of the speech signal can be written as

$$|X()| = |E()|\ \ |V()| \qquad (1)$$

Taking the logarithm of above equation yields

$$\log|X()|= \log|E()| + \log|V()|. \qquad (2)$$

Particularly for voiced sounds, it can be observed that the E term corresponds to an event that is relatively extended in time (e.g., a pulse train with pulses every 10 ms), and thus it yields a spectrum that should be characterized by a relatively rapidly varying function; in comparison, because of the relatively short impulse response of the vocal tract, the V term varies more slowly with function. With the use of this knowledge, the left-hand side of equation(2) can be separated into the two right-hand-side components by a kind of a filter that separates the log spectral components that vary slowly with function(the so called high-time components) from those that vary slowly with function (the low-time components). Such an operation would essentially be performing deconvolution.

Equation (2) has transformed the multiplicative formula (1) into a linear operation and thus can be subjected to linear operations such as filtering. Since the variable is frequency rather than time, notations must be changed. Thus, rather than filtering (for time), filter (for frequency); instead of a frequency response, use quefrency response; and the DFT (or z transform or Fourier transform) of the log |X()| is called the cepstrum. The cepstrum is computed by taking the inverse z transform of equation 2 on the unit circle, yielding

where : c(n) is called the nth cepstral coefficient.

III. HIDDEN MARKOV MODEL

In speech recognition, the basic idea is to find the most likely string of words given some acoustic input, or: arg max P(w / y)

where w is a string of words

y is the set of acoustic vectors that comes from the cepstrum output.

The acoustics are observations, and the words are sequences. Words are made of ordered sequences of phonemes: /h/ is followed by /e/ and then by /l/ in the word \hello". Each phoneme can in turn be considered as a particular random process (possibly Gaussian). This structure can be adequately modeled by a left-right HMM, where each state correspond to a phone. In \real world" speech recognition, the phonemes themselves are often modeled as left-right HMM's rather than plain Gaussian densities (e.g. to model separately the attack, then the stable part of the phoneme and finally the \end" of it). Words are then represented by large HMM's made of concatenations of smaller phonetic HMM's.

In speech recognition, it is useful to associate an \optimal" sequence of states to a sequence of observations, given the parameters of a model. For instance, in the case of speech recognition, knowing which frames of features \belong" to which state allows to locate the word boundaries across time. This is called the alignment of acoustic feature sequences. A “reasonable" optimality criterion consists in choosing the state sequence (or path) that brings a maximum likelihood with respect to a given model. This sequence can be determined recursively via the Viterbi algorithm.

This algorithm makes use of two variables: a) $\delta t(i)$ is the highest likelihood along a single path among all the paths ending in state i at time t : b) a variable $\psi t(i)$ which allows to keep track of the “best path" ending in state i at time t :. The idea of the Viterbi algorithm is to find the most probable path for each intermediate and terminating state in the trellis Only this most likely path 'survives' [11].

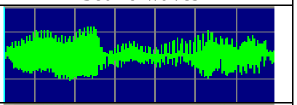
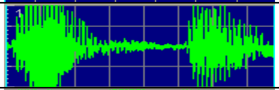
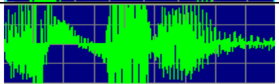
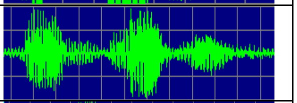
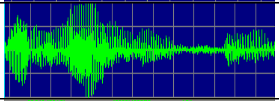
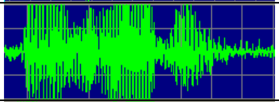
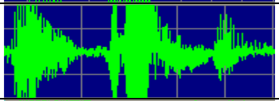
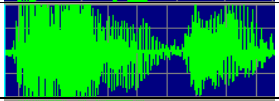
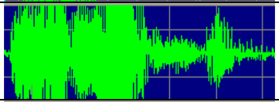
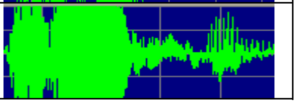
IV. RADIAL BASIS FUNCTION (RBF)

RBF is a supervised ANN which works based on the distance concept. The distance is found between a pattern and each centre. The centre is also one of the patterns predefined. The square of the distance is a node in the hidden layer. An exponential function is used as an activation function which will be the output of the particular node. The number of nodes in the hidden layer is based on the number of centres decided in an implementation [12].

A. *EXPERIMENTAL SETUP*

The wave files were collected using a microphone samples at 8khz. Ten different words collected are presented in. The words (Table 1) are the speech uttered by both men and woman.

TABLE I. WORDS AND THEIR SPEECH WAVES

S.No	Words	Sound waves
1	Fourier	
2	Transform	
3	Filtering	
4	Deblurring	
5	Enhancing	
6	Coloring	
7	Compression	
8	Fanbeem	
9	Dilation	
10	Erosion	

B. *IMPLEMENTATION*

The following steps are adapted to train and test the words given in Table 1.

Cepstral coefficients

1) *Acquire wave file.*
2) *Remove zeros which does not give any information.*
3) *Apply linear predictive analysis.*
4) *Apply fast fourier transform.*
5) *Apply log for the output in step 4.*
6) *Apply inverse fast fourier transform.*
7) *Apply levinson Durban equation.*
8) *Repeat step 3 to step 7 for every 10 samples of the data acquired from wave files average all values to finally get only 10 values.*
9) *Repeat step 8 for the data collected from the other 10 words. Hence after averaging there will be 10 patterns with each pattern having 10 values.*
10) *Repeat step 9 for other words.*

C. *HMM training*

Build each HMM by using cepstral feature vectors to produce cluster sets (Clusterer), initializing the model with appropriate values based on the cluster set (Initializer) and then improving the model parameters to best represent the original data (Trainer). Once an HMM for each word is obtained, make use of the model in applications: from training data so as to test the model's accuracy by (Tester). During the process of HMM training the viterbi algorithm uses Initialization, Recursion, Termination and backtracking

D. *Radial basis function*

Each HMM model of a word is associated with a target value. Similarly we have 10 HMM model and hence ten target values. For each value,

1) *Decide number of centres*
2) *Find the distance between a pattern and a centre and hence for all centres. The number of centres is based on the number of nodes in the hidden layer*
3) *Apply exponential (-x) activation and hence RBF is obtained.*
4) *Add a bias of '1' to the RBF and hence obtain an RBF matrix (10 X 4). The value 10 represents 10 words and 4 represents (3 RBF and bias '1').*
5) *Find the inverse of the RBF matrix and process with the target values to obtain the final weight.*

V. RESULTS AND DISCUSSIONS

The speech was processed by cepstrum method to obtain feature vectors. The feature vectors were given for HMM formulation.

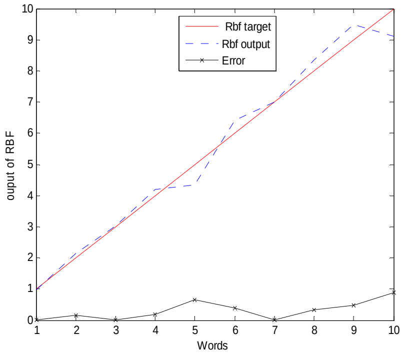


Figure 2. Output RBF for all the

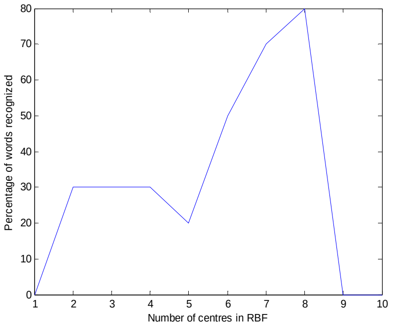


Figure 3. Effect of number of centres in recognizing the words

The output of HMM was given as input for RBF network. Target values were fixed to train the RBF. The training was carried out with different number of centres (Figure 2). When 8 centres were used, the percentage of word recognition resulted in 80%. The output of RBF is given in Fig. 1.

VI. CONCLUSIONS

In this research work, 10 words have been considered. Cepstrum analysis has been used to extract features of vectors and HMM model was developed for all the ten words. Target values were fixed for all the 10 words and trained with RBF. When the speech was tested with HMM model and RBF, the proposed approach gives 80% performance. That is only 8 words were recognized. As a future work, the number of centres have to be changed to find out improvement in the performance of the speech recognition.

REFERENCES

[1] Y. Gong, “Speech recognition in noisy environments: A survey,” Speech Communication, vol. 16, pp. 261–291, 1995.

[2] S. F. Boll, “Suppression of acoustic noise in speech using spectral subtraction,” IEEE Trans. on Acoustics, Speech, and Signal Processing, vol. ASSP-27, no. 2, pp. 113–120, 1979.

[3] M. J. F. Gales and S. J. Young, “Robust continuous speech recognition using parallel model combination,” IEEE Trans.on Speech, and Audio Processing, vol. 4, pp. 352–359, 1996.

[4] M. Cooke, P. Green, L. Josifovski, and A. Vizinho, “Robust automatic speech recognition with missing and unreliable acoustic data,” Speech Communication, vol. 34, pp. 267–285, 2001.

[5] J. P. Barker, M. P. Cooke, and D. P. W. Ellis, “Decoding speech in the presence of other sources,” Speech Communication, vol. 45, pp. 5–25, 2005.

[6] A. Drygajlo and M. El-Maliki, “Speaker verification in noisy environment with combined spectral subtraction and missing data theory,” in Proc. ICASSP ’98, 1998, vol. 1, pp. 121–124.

[7] M. L. Seltzer, B. Raj, and R. M. Stern, “Classifier based mask estimation for missing feature methods of robust speech recognition,” in Proc. ICSLP ’00, 2000, pp. 538–541.

[8] G. J. Brown, J. Barker, and D. L. Wang, “A neural oscillator sound separator for missing data speech recognition,” in Proc. IJCNN ’01, 2001, pp. 2907–2912.

[9] H. Van Hamme, “Robust speech recognition using cepstral domain missing data techniques and noisy masks,” in Proc.ICASSP ’04, 2004, vol. 1, pp. 213–216.

[10] A. S. Bregman, “Auditory scene analysis”, The MIT Press, Cambridge, MA, 1990.

[11] Rabiner, L. R. (1989), “A tutorial on hidden Markov models and selected applications in speech recognition”, Proceedings of the IEEE, 77 (2), 257-286.

[12] P. Raviram, R. S. D. Wahidabanu, and S. Purushothaman, “Concurrency Control in CAD With KBMS Using Counter Propagation Neural Network”, IEEE International Advanced Computing Conference, March 6-7, 2009, Thapar University, Patiala, India

[13] Mohamad Adnan Al-Alaoui, Lina Al-Kanj, Jimmy Azar, and Elias Yaacoub, ”Speech Recognition using Artificial Neural Networks and Hidden Markov Models” , IMCL2008 Conference 16-18 April 2008.

[14] J. Azar, H. Abou Saleh, and M. A. Al-Alaoui, “Sound Visualization for the Hearing Impaired,” International Journal of Emerging Technologies in Learning" - iJET. March, 2007, pp. 1-7.

[15] Ferzli R. and M. A. Al-Alaoui, "Subsampling Image Compression Using Al-Alaoui Backpropagation Algorithm", the 14th IEEE International Conference on Electronics, Circuits and Systems, Marrakech, Morocco, December 11-14, 2007.

[16] A. Pinkus, “Approximation theory of the NLP model in neural networks", tech. rep., 1999.

[17] A. Sarkar and T.V. Sreenivas, “Automatic Speech Segmentation Using Average Level Crossing Rate Information”.

[18] S. Al Hattab, Y. Yaacoub, A. El hajj, “Teaching and Learning Using Information Technology Arabic Speech Recognition Team”, FYP, ECE, AUB, May 2007.