

New Feature Extraction Techniques for Marathi Digit Recognition

N S. Nehe¹ and R. S. Holambe²

S.G.G.S. Institute of Engineering and Technology, Nanded, (MS-INDIA)

¹nsnehe@yahoo.com ²rsholambe@sngs.ac.in

Abstract— In this paper a new efficient feature extraction methods for speech recognition have been proposed. The features are obtained from Cepstral Mean Normalized reduced order Linear Predictive Coding (LPC) coefficients derived from the speech frames decomposed using Discrete Wavelet Transform (DWT). In the literature it is assumed that the speech frame of size 10 msec to 30 msec is stationary, however, in practice different parts of the speech signal may convey different amount of information (hence may not be perfectly stationary). LPC coefficients derived from wavelet-decomposed subbands of speech frame provide better representation than modeling the frame directly. Experimentally it has been shown that, the proposed approach provides effective (better recognition rate), efficient (reduced feature vector dimension) features. The speech recognition system using the Continuous Density Hidden Markov Model (CDHMM) has been implemented. The proposed algorithms were evaluated using isolated Marathi digits database in presence of white Gaussian noise.

Index Terms—Feature Extraction, Linear Predictive Coding, Discrete Wavelet Transform, Cepstral Mean Normalization.

I. INTRODUCTION

All speech recognizers include an initial signal processing front end that converts a speech signal into its more convenient and compressed form called feature vectors. Feature extraction method plays a vital role in speech recognition task. There are two dominant approaches of acoustic measurement. First is in temporal domain approach (parametric) like Linear Prediction [1], which is developed to closely match the resonant structure of human vocal tract that produces the corresponding sound. Second is frequency domain approach (nonparametric) known as Mel-Frequency Cepstral Coefficients (MFCC)[2]. In another approach [3,4] wavelet transform and wavelet packet tree have been used for speech feature extraction in which the energies of wavelet decomposed subbands have been used in place of Mel filtered subband energies. However, the time information is lost due to use of wavelet subband energies. In our approach we use the actual wavelet coefficients, which preserve the time information.

We propose new feature extraction methods in which each frame of speech signal is decomposed into different

frequency subbands using Discrete Wavelet Transform (DWT) and each subband is further modeled using Linear Predictive Coding (LPC). Further the Cepstral Mean Normalization (CMN) of obtained feature vectors gives the new effective and robust features with small dimension abbreviated as Wavelet sub-band Cepstral Mean Normalized features (WSCMN). In the literature it is assumed that the speech frame of size 10 msec to 30 msec is stationary [5]. However, different parts of the frame may contain different information. Hence, in this paper an attempt has been made to derive effective, efficient and robust features from the frequency subbands of the frame. The subband decomposition has been obtained by means of DWT. DWT is more popular in the field of Digital Signal Processing (DSP) due to its multiresolution capability. Also it has the property of constant Q, which is one of the demands of many signal processing applications, especially in the processing of the speech signals (as human's hearing system is constant Q perceptual) [6]. Wavelet decomposition results in a logarithmic set of bandwidths, which is very similar to the response of human ear to frequencies (logarithmic fashion). We thus use DWT for subband decomposition.

II. DISCRETE WAVELET TRANSFORM AND WSCMN

The continuous wavelet transform (CWT) is given by (1) where the function $\psi(t)$, j and k are called the (mother) wavelet, scaling factor and shift parameter respectively.

$$W(j, k) = \frac{1}{\sqrt{j}} \int_{-\infty}^{\infty} x(t) \psi^* \left(\frac{t-k}{j} \right) dt \quad (1)$$

As CWT is a function of two parameters, it contains high redundancy while analyzing the functions. Instead of this if we analyze the signal with small number of scales with varying number of translations at each scale gives DWT. DWT theory [7] requires two sets of related functions called scaling function and wavelet function given by

$$\phi(t) = \sum_{n=0}^{N-1} h[n] \sqrt{2} \phi(2t - n) \quad (2)$$

and

$$\psi(t) = \sum_{n=0}^{N-1} g[n] \sqrt{2} \phi(2t - n) \quad (3)$$

where, function $\phi(t)$ is called scaling function, $h[n]$ is an

N. S. Nehe is a research scholar with the S.G.G.S. Institute of Engineering and Technology, Nanded (INDIA) (E-mail: nsnehe@yahoo.com). (Corresponding Author)

Dr. R. S. Holambe, Professor, is with the Instrumentation Engineering Department, S.G.G.S. Institute of Engineering and Technology, Nanded (INDIA) (E-mail: rsholambe@sngs.ac.in).

impulse response of a low pass filter and $g[n]$ is an impulse response of a high pass filter.

In speech signal, high frequencies are present very briefly at the onset of a sound while lower frequencies are present latter for long period [7]. DWT resolve all these frequencies well. The DWT parameters contain the information of different frequency scales. This helps in getting the speech information of corresponding frequency band. We thus introduce a DWT to decompose speech signal into the frequency bands. In order to parameterize the speech signal, we first decompose the signal into four frequency bands in uniform/dyadic fashion. Most of the information in the human speech is only contained in a few scales of Wavelet Transform (WT) decomposition [6]. Hence a few (2-3) number of WT scales help to reduce the final feature vector dimensions.

Figure 1 shows the block diagram of proposed feature extraction methods. In the first method, two level uniform decomposition of preprocessed and windowed speech frame has been done using wavelet packets (Figure 1(a)). Here 32nd order Daubechies's wavelet is used. In the second method three levels DWT decomposition of preprocessed and windowed speech frames has been done using Daubechies's wavelet filters (Figure 1(b)). Dyadic decomposition mimics the human perception system hence decomposition is done in dyadic fashion. Actual wavelet coefficients retain the time information [4], hence LPC features have been estimated from the DWT coefficients in time domain. LPC features of p th order have been extracted from each subband of wavelet decomposed speech signal (in both the methods). These sub-vectors of LPC coefficients obtained from each subband are concatenated to form a final LPC feature vector. Thus the LPC feature vector f_i derived from frame i for uniform and dyadic method can be expressed by equation (4) and (5) respectively.

$$f_i = [a_{AA} \quad a_{AD} \quad a_{DA} \quad a_{DD}]^T, \quad (4)$$

$$f_i = [a_{A_3} \quad a_{D_3} \quad a_{D_2} \quad a_{D_1}]^T, \quad (5)$$

where, a_{A_j} is a row vector formed using LPC coefficients obtained from A_j and a_{D_j} is row vector formed using prediction coefficients obtained from D_j ($j = 1,2,3$).

Similarly, $a_{AA}, a_{AD}, a_{DA}, a_{DD}$ are the LPC coefficients obtained from the uniform decomposed subbands (refer Figure 1(a)). T indicates a vector transpose. LPC cepstrums (LPCC) were then obtained from LPC feature vectors (f_i). Finally CMN [8] of LPCC were done to obtain the proposed WSCMN features.

Cepstral Mean Normalization (CMN) is the simplest feature normalization technique to implement. It provides many of the benefits available in the more-advanced normalization algorithms. The received speech signal $x(n)$ is used to estimate a sequence of LPCC vectors $\{x_1, x_2, \dots, x_T\}$. In its basic form, CMN consists

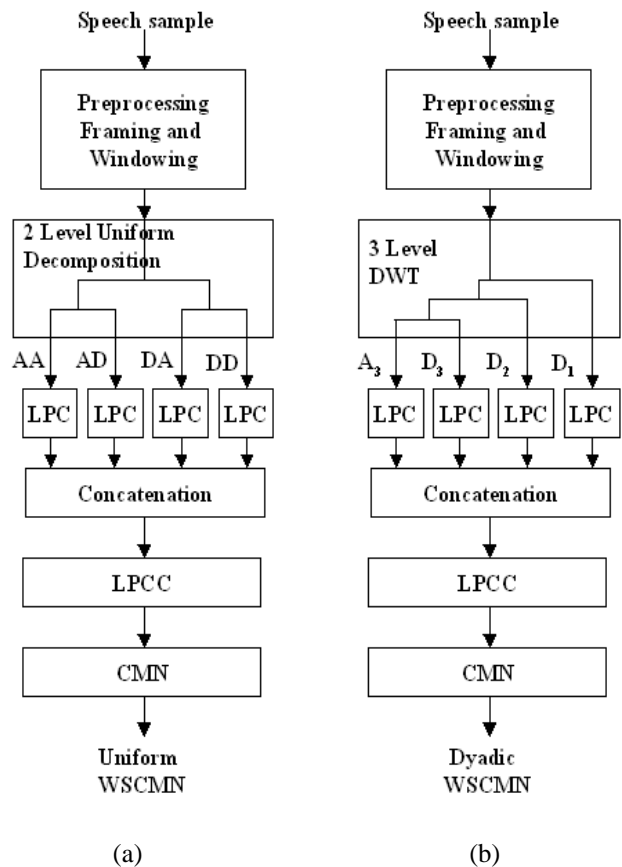


Figure 1. WSCMN feature extraction steps (a) Uniform Decomposition (b) Dyadic Decomposition

of subtracting the mean feature vector μ_x from each vector x_t and normalizing by variance σ_x to obtain the normalized (WSCMN) vector \hat{x}_t .

$$\hat{x}_t = \frac{x_t - \mu_x}{\sigma_x}, \quad (6)$$

$$\mu_x = \frac{1}{T} \sum_t x_t \quad \text{and} \quad \sigma_x^2 = \frac{1}{T} \sum_{t=1}^T x_t^2 - \mu_x^2 \quad (7)$$

After normalization, the mean of the cepstral sequence is zero, and it has a variance of one. CMN makes the features robust to some linear filtering of the acoustic signal, which might be caused by microphones with different transfer functions, varying distance from user to microphone, the room acoustics, or transmission channels.

III. EXPERIMENTS AND RESULTS

The database used for experimentation was ZERO to NINE digits in Marathi language. The data was collected from five male and five female speakers (from students of our institute). Each word was uttered 20 times by each speaker and recorded with sampling frequency 10 kHz. Out of 20 utterances per word per speaker ten were used for training and remaining ten were used for testing. Hence, the complete data size was 1000 samples for training and 1000 samples for testing.

The silence part of speech signal was removed manually and the signal was preemphasized using first-order filter with impulse response $[1 \ -0.97]$. Then the signal was divided into overlapping frames of size 32 msec each (320 samples for 10 kHz sampling frequency) with 50% overlap. For windowing, Hamming window was used.

In the first type, MFCC using triangular Mel-filters were estimated from each frame [2]. Frame spectrum was warped through a triangular Mel filter bank. Twenty Mel-filters were used in a filter bank. Energy of each filter output was estimated. Cepstral coefficients were estimated by applying Discrete Cosine Transform (DCT) on log energies. MFCC feature vector of length 39 was constructed from first 13 MFCC coefficients and their first and second derivatives. The recognition rate obtained using MFCC has been used as a baseline for comparison. In the second method, MFCC features were normalized using CMN and used for comparison. These features are denoted as MFCC-CMN features.

In the third type, speech frame was decomposed into subbands of uniform bandwidth by two level wavelet packet transform [7]. Daubechies's wavelet with 32nd order was used for decomposition. Prediction coefficients with 5th order LPC (as it gives the best performance) were estimated from the subbands. The prediction coefficients estimated from the subbands were then concatenated to form f_i vector (for i th frame). Finally CMN of cepstrums obtained from f_i were estimated to get the proposed uniform WSCMN (U-WSCMN) features.

Finally in the last method, instead of uniform decomposition dyadic decomposition of speech frames was done and the dyadic WSCMN features (D-WSCMN) were obtained similar to uniform WSCMN.

Five prediction coefficients from each subband (in third and fourth method) give feature vector of dimension 20. Performances of these features were tested using Left-Right continuous HMM with 3 mixtures and 4 states (as this combination yields best performance) on clean database of Marathi digits. Also the robustness of the proposed features was tested on the noisy data obtained by adding the white Gaussian noise for various Signal to Noise Ratio (SNR) in clean data samples. Table 1 shows the performance of features in terms of % recognition rate in clean as well as noisy conditions.

TABLE I. PERFORMANCE OF FEATURES AT VARIOUS NOISE LEVELS

SNR dB	MFCC	MFCC-CMN	Uniform WSCMN	Dyadic WSCMN
Clean	85.40	79.80	100.00	100.00
30	83.30	74.90	98.70	99.00
20	74.20	66.70	92.80	96.90
15	56.60	61.80	75.70	82.90
10	43.30	48.70	49.90	54.40
5	32.10	42.10	30.00	45.60
0	20.30	23.40	22.90	26.50

Figure 2 shows the average recognition performance of various features in presence of additive white Gaussian

noise. This shows that the proposed features perform better over MFCC and MFCC-CMN.

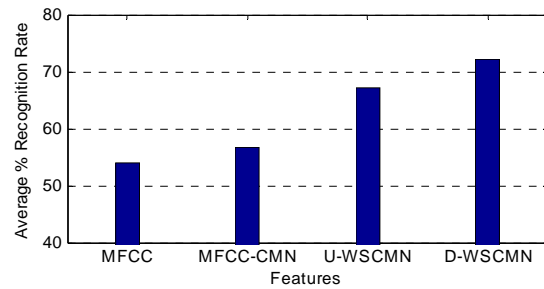


Figure 2: Average recognition performance of features

IV. CONCLUSIONS

This paper has presented new effective and robust feature extraction methods using DWT and LPC for isolated Marathi digits recognition. Experimental results show that the proposed WSCMN features yield better performance over MFCC and CMN and also give 100% recognition performance on clean data. Feature dimension for WSCMN is almost half of the MFCC. This reduces the memory requirement and the computational time. It is also observed that the performance of dyadic WSCMN is better than uniform WSCMN in noisy recognition because the dyadic (logarithmic) frequency decomposition mimics the human auditory perception system more better than uniform frequency decomposition. So the proposed approach provides effective (better recognition rate), efficient (reduced feature vector dimension) and robust features. Hence these features can be used for real time speech recognition using DSP processors.

REFERENCES

- [1] F. Itakura, "Minimum prediction residual principle applied to speech recognition," *IEEE Trans. Acoust. Speech, Signal Proces.*, vol. ASSP-23, pp. , pp. 67-72, 1975.
- [2] S. B. Devis and P. Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *IEEE Trans. Acoust., Speech, and Signal Proces.*, vol. ASSP-28, no. 4, August 1980.
- [3] Z. Tufekci and J. N. Gowdy, "Feature extraction using discrete wavelet transform for speech recognition," *IEEE Inter. Conf. Southeastcon2000*, pp. 116-123, April 2000.
- [4] O. Farooq and S. Datta, "Mel filter-like admissible wavelet packet structure for speech recognition," *IEEE Signal Process. Letters*, vol. 8, no. 7, July 2001.
- [5] L. R. Rabiner and B. H. Juang, *Fundamentals of Speech Recognition*. Englewood Cliffs, NJ: Prentice-Hall Inc. 1993.
- [6] Y. Hao and X. Zhu, "A new feature in speech recognition based on wavelet transform," *In Proc. IEEE 5th Inter. Conf. on Signal Processing (WCCC-ICSP 2000)*, vol. 3 pp. 1526-1529, 21-25 August 2000.
- [7] K. P. Soman and K. I. Ramchandran, *Insight into Wavelets from Theory to Practice*. Prentice-Hall of India, 2e. 2000.
- [8] S. Furui, "Cepstral analysis technique for automatic speaker verification," *IEEE Trans. ASSP-29*, pp. 256-272, 1981.