Robust Speech Recognition Using Wavelet Domain Front End and Hidden Markov Models

Rajeswari, N. N. S. S. R. K. Prasad and V. Sathyanarayana

Abstract This paper presents a method to address the issue of noise robustness using wavelet domain in the front end of an automatic speech recognition (ASR) system, which combines speech enhancement and the feature extraction. The proposed method includes a time-adapted hybrid wavelet domain speech enhancement using Teager energy operators (TEO) and dynamic perceptual wavelet packet (PWP) features applied to a hidden Markov model (HMM)–based classifier. The experiments are performed using the HTK toolkit for speaker-independent database which are trained in a clean environment and later tested in the presence of AWGN. It has been seen from the experimental results that the proposed method has a better recognition rate than the most popular MFCC-based feature vectors and HMM-based ASR in noisy environment.

Keywords Automatic speech recognition · Perceptual wavelet packet · Hidden Markov model · Mel-frequency cepstral coefficients · Teager energy operator · Additive white gaussian noise

Rajeswari (✉)
Department of Electronics and Communication Enginering, Acharya Institute of Technology, Bangalore, India
e-mail: rajeswari@acharya.ac.in

N. N. S. S. R. K.Prasad
ADA, Ministry of Defence, Goverment of India, Bangalore, India
e-mail: nnssrkprasad2007@gmail.com

V. Sathyanarayana
Department of DSP, CET, Jain University, Bangalore, India
e-mail: satyaec49@gmail.com

V. Sridhar et al. (eds.), *Emerging Research in Electronics, Computer Science and Technology*, Lecture Notes in Electrical Engineering 248,
DOI: 10.1007/978-81-322-1157-0_44,

1 Introduction

An ASR system finds large applications demanding for real-time environments which are embedded with high ambient noise levels. The performance of an ASR is acceptable in clean environments; however, the system performance degrades in the presence of noise. Thus, there is a strong need for noise robustness to be considered [1–3].

Much of the research in speech recognition aims at first robust feature extraction and the second being building a robust classifier [4]. The most popular MFCC features based on short-time Fourier transform (STFT) and power spectrum estimation do not give a good representation of noisy speech. The features based on the STFT produce uniform resolution over the time–frequency plane. Due to this, it is difficult to detect sudden bursts in a slowly varying signal or the highly non-stationary parts of the speech signal. The recent approach of wavelet packets which segment the frequency axis and makes uniform translation in time is been proposed. Wavelet coefficients provide flexible and efficient manipulation of a speech signal localized in the time–frequency plane which is an alternative to MFCC [5–7]. The perceptual wavelet filter bank is built to approximate the critical band responses of the human ear. Wavelet packets decompose the data evenly into all bins, but PWPs decompose only critical bins [8].

In this paper, we propose a wavelet domain front end for an ASR which combines speech enhancement and the feature extraction. The proposed method includes a time-adapted wavelet domain hybrid speech enhancement using Teager energy operators (TEO) and dynamic perceptual wavelet packet (PWP) features applied to a hidden Markov model (HMM)–based classifier.

The rest of the paper is organized as follows. Section 2 introduces a block diagram of the proposed wavelet domain front end of the ASR and provides detailed description of each constituting part. Section 3 describes the recognizer and the toolkit used. Section 4 evaluates the performance of the proposed system under different levels of noise. The conclusion is presented in Sect. 5.

2 Proposed Wavelet Domain Front End of ASR

Figure 1 describes the proposed noise robust wavelet domain front end of an ASR. The noisy speech input wave files are sampled at 16 kHz and segmented into frames of 24-ms duration with frame shift interval of 10-ms overlap.

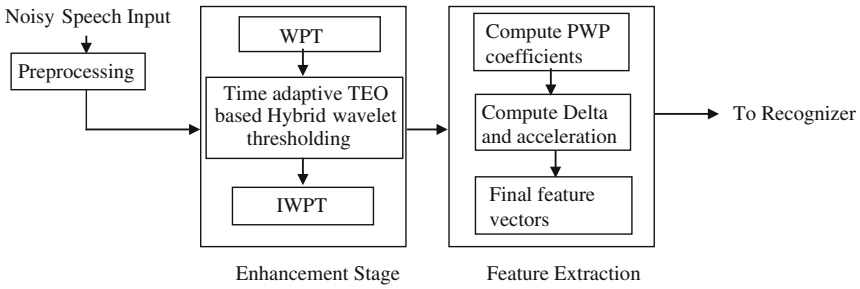


Fig. 1 Block diagram of proposed wavelet domain front end of ASR

2.1 Speech Enhancement

In the proposed method, wavelet packet transform (WPT) is applied to each input frame. The coefficients obtained are then subjected to Teager energy approximation [9, 10], where the threshold is adapted with respect to the voiced/unvoiced segments of the speech data. A hybrid thresholding process is adopted which is a compromise for the conventional hard and soft thresholding in preserving both the edges and reducing the noise.

2.1.1 Wavelet Packet analysis

For a j-level WP transform, the noisy speech signal $y[n]$ with frame length N is decomposed into 2^j sub-bands. The mth WP coefficient of the kth sub-band is expressed as follows:

$$W_{k,m}^{J} = \mathrm{WPT}\{y(n), j\} \tag{1}$$

where $n = 1,\ldots N$, $m = 1,\ldots, N/2^j$ and $k = 1,\ldots 2^j$.

2.1.2 Teager Energy Operator on Wavelet Coefficients

– Teager energy approximation for each WPT sub-band is computed

$$\mathrm{TEO_{i,k}} = \mathrm{Y_{i,k}^2} - \mathrm{Y_{i,k-1}Y_{i,k+1}} \tag{2}$$

– TEO coefficients are smoothened in order to reduce the sensitivity to noise

$$M_{i,k} = \mathrm{TEO}_{i,k} * H_p \tag{3}$$

– Normalize the TEO coefficients

$$M'_{i,k} = \left[\frac{M_{i,k}}{\max\left(M_{i,k}\right)} \right] \tag{4}$$

– Time-scale adaptive threshold based on Bayes shrink for each sub-band k is computed.

$$\lambda_{i,k} = \lambda_i (1 - M'_{i,k}) \tag{5}$$

2.1.3 Denoising by Thresholding

Denoising using wavelet packet coefficients is performed by thresholding; that is, the coefficients which fall below the specific value are shrunk and the later retained. Different thresholding techniques have been proposed. However, there are two popular thresholding functions used in the speech enhancement systems which are the hard and the soft thresholding functions [11–13].

Hard thresholding is given by

$$T_s(\lambda, w_k) = \begin{cases} w_k & \text{if } |w_k| > \lambda \\ 0 & \text{if } |w_k| \leq \lambda \end{cases} \tag{6}$$

Soft thresholding is given by

$$T_s(\lambda, w_k) = \begin{cases} \text{sgn}(w_k)|w_k| - \lambda & \text{if } |w_k| > \lambda \\ 0 & \text{if } |w_k| \leq \lambda \end{cases} \tag{7}$$

where w_k represents wavelet coefficients and λ the threshold value.

However, hard thresholding is best in preserving edges but worst in denoising, while soft thresholding is best in reducing noise but worst in preserving edges. In order to have a general case of both reducing noise and preserving edges, a hybrid thresholding is used.

Hybrid thresholding is given as follows:

$$T_s(\lambda, w_k) = \begin{cases} w_k * \frac{|w_k|^\alpha - \lambda^\alpha}{|w_k|^\alpha} & \text{if } |w_k| > \lambda \\ 0 & \text{if } |w_k| \leq \lambda \end{cases} \tag{8}$$

With careful tuning of parameter α for a particular signal, one can achieve best denoising effect within thresholding framework.

The enhanced speech is then reconstructed using the inverse WP transform

$$x'(n) = \text{WPT}^{-1}\left\{W'^{J}_{K}, j\right\} \tag{9}$$

2.2 Dynamic Perceptual Wavelet Packet Feature Extraction

Wavelet coefficients provide flexible and efficient manipulation of a speech signal localized in the time–frequency plane [5–7]. The perceptual wavelet filter bank is built to approximate the critical band responses of the human ear. Wavelet packets decompose the data evenly into all bins but PWPs decompose only critical bins [5, 8, 14]. The size of the decomposition tree is directly related to the number of critical bins. The decomposition is implemented by an efficient 7-level tree structure, depicted in Fig. 2. The PWP transform is used to decompose $nx(n)$ into several frequency bands that approximate the critical bands. The terminal nodes of the tree represent a non-uniform filter bank.

The PWP coefficients for the sub-bands are generated as follows:

$$wj, i(k) = \text{pwpt}\,(nx\,(n)) \tag{10}$$

where $n = 1, 2, 3, \ldots L$ (L is the frame length),

$j = 0, 1, 2, \ldots, 7$ (j is the number of levels),
$i = 1, 2, 3, \ldots (2^j - 1)$ (i is the sub-band index in each level of j).

The static PWP coefficients are made more robust by computing the delta and the acceleration coefficients.

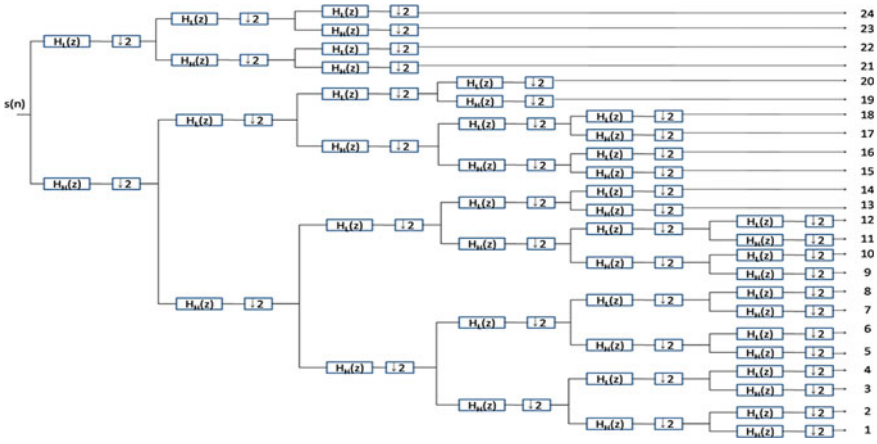


Fig. 2 Tree structure of PWPT

3 Speech Recognition

HMMs are the most successful statistical models for classification of speech. HMM is a stochastic approach which models the given problem as 'doubly stochastic process' [15–17].

A N-state HMM is defined by the parameter set $\lambda = \{\pi_i, a_{ij}, b_i(x), i,j = 1,\ldots N\}$, where

π_i initial state probability for state i,

a_{ij} transition probability from state i to state j,

$b_i(x)$ state observation probability density function (pdf) that is usually modeled by a mixture of Gaussian densities.

A five-state continuous density HMM is used as the statistical model for classification of speech signals. The hidden Markov model toolkit (HTK) is a portable toolkit for building and manipulating HMMs optimized for speech recognition [18].

4 Experimental Setup and Results

To evaluate the performance of the proposed method, recognition experiments were carried out using TIMIT database. The data were digitized with sampling rate of 16 kHz and 16 bits/sample quantization. This database consists of 200 training sentences from 6 male and female speakers and testing sentences randomly picked from training data. To simulate various noisy conditions, the testing sentences were corrupted by the additive white Gaussian noise with SNR conditions from 40 to 0 dB. The baseline recognition system was implemented on HTK toolkit with continuous density HMM models.

In the proposed feature extraction method, the perceptual wavelet features are extracted using MATLAB and written into HTK format using htkwrite function available from voicebox MATLAB package. The 24 perceptual wavelet filter banks are constructed by trial and error. The proposed tree shows excellent results using the Daubechies 45 prototype. Using a hamming windowed analysis frame length of 20 ms, the resulted 13-dimensional features plus their delta and delta–delta features, in other word, totally 39-dimensional features were used for speech recognition.

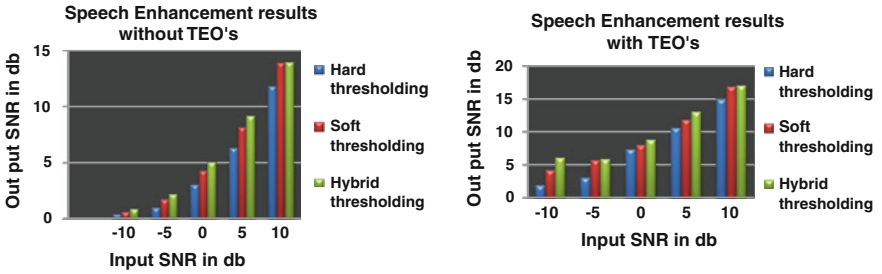
Table 1 shows that compared to the standard wavelet thresholding method, time-adapted wavelet-based hybrid thresholding using the TEOs as proposed in this paper has a better enhancement effect for SNR levels ranging from −10 to +10 db. The proposed wavelet domain front end features with HMM classifier–based ASR performs better recognition when compared to the conventional ASR with MFCC features and HMM as shown in Table 2. The PWPs capable of decomposing the critical bins during feature extraction supports in providing robust features for the ASR.

Table 1 Input/output SNR values obtained with different thresholding methods

Input SNR/output SNR (in db)		−10	−5	0	5	10
Hard thresholding	Without TEO	0.4184	1.2036	3.0537	6.2870	11.7655
	With TEO	1.9318	3.0627	7.3341	10.6041	14.9781
Soft thresholding	Without TEO	0.6155	1.7538	4.2496	8.1132	13.8283
	With TEO	4.2299	5.7624	8.0517	11.8721	16.8563
Hybrid thresholding	Without TEO	0.8724	2.1756	5.0328	9.1108	13.8886
	With TEO	6.0823	5.9035	8.8096	13.0061	17.0388

Table 2 Comparison of recognition accuracy for different features at various noise levels

Noise level	MFCC_E_D_A	MWAVELET_E_D_A
0 db	20.33	21
5 db	41.23	49.36
10 db	68.66	71.26
20 db	79.88	82.46
25 db	82.63	86.43
30 db	85.55	89.65
35 db	89.66	90.35
40 db	91.56	93.24



5 Conclusions

A wavelet domain front end for an ASR which combines both the enhancement and the feature extraction for different noise levels has been presented. The proposed time-adapted wavelet-based hybrid thresholding using the TEOs outperforms the conventional wavelet-based denoising schemes, as shown in Table 1. The dynamic PWP features, which decomposes only the critical sub-bands applied to a HMM–based classifier along with the proposed enhancement algorithm, recognize better than the conventional MFCC features with HMM as shown in Table 2. Further the ASR can be improved with a hybrid classifier into context and for larger database.

Acknowledgments We would like to express our sincere thanks to Aeronautical Development Agency, Ministry of Defence, DRDO, Bangalore, India, for supporting to do our research work.

References

1. Gong Y (1995) Speech recognition in noisy environments: a survey. Speech Commun 16(3):261–291
2. Rabiner L, Juang BH (1996) Fundamentals of speech recognition, vol 103 Prentice Hall Englewood Cliffs, New Jersey
3. O'Shaughnessy D (2001) Speech communication: human and machine. IEEE Press
4. Yusnita MA (2011) Phoneme-based or isolated-word modelling speech recognition system. In: IEEE 7th international colloquium on signal processing and its applications, pp 304–309
5. Jiang H et al (2003) Feature extraction using wavelet packet strategy. In: Proceedings of 42nd IEEE conference on decision and control, pp 4517–4520
6. Jie Y, Zhenli W (2009) Noise robust speech recognition by combining speech enhancement in the wavelet domain and Lin-log RASTA. ISECS 415–418
7. Gupta M, Gilbert A (2002) Robust speech recognition using wavelet coefficient features. IEEE Trans Speech Audio Process 445–448
8. Bourouba H et al (2008) Robust speech recognition using perceptual wavelet de-nosing and Mel-frequency product spectrum Cepstral coefficient features. Proc Informatica 32:283–288
9. Hesham T (2004) A time-space adapted wavelet de-noising algorithm for robust ASR in low SNR environments. IEEE Trans Speech Audio Process 1:311–314
10. Bahoura M, Rouat J (2001) Wavelet speech enhancement based on the Teager energy operator. Signal Process Lett IEEE 8(1):10–12
11. Chang S, Yu B, Vetterli M (2000) Adaptive wavelet thresholding for image denoising and compression. IEEE Trans Image Process 9(9):1532–1546
12. Donoho DL, Johnstone IM (1995) De-noising by soft-thresholding. IEEE Trans Inf Theory 41(3):613–627
13. Donoho DM, Johnstone IM (1995) Adapting to unknown smoothness via wavelet shrinkage. J Am Stat Assoc 90(432):1200–1224
14. Mallat S (2001) A wavelet tour of signal processing. Academic Press, London
15. Jisn H et al (2006) Large margin HMMs for speech recognition. IEEE Trans Speech Audio Lang Process 14(5):1584–1595
16. Vaseghi SV, Milner BP (1997) Noise compensation methods for HMM speech recognition in adverse environments. IEEE Trans Speech Audio Process 5(1):11–21
17. Mark J, Gales F, Young SJ (1996) Robust continuous speech recognition using parallel model combination. IEEE Trans Speech Audio Process 4(5):352–359
18. Young S (2009) The HTK book. Version 3.4. Cambridge University Engineering Department. Cambridge, UK