Engineering Science and Technology, an International Journal 17 (2014) 145–151

HOSTED BY

ELSEVIER

Contents lists available at ScienceDirect

Engineering Science and Technology, an International Journal

journal homepage: http://www.elsevier.com/locate/jestch

Full length article

Auditory ERB like admissible wavelet packet features for TIMIT phoneme recognition

P.K. Sahu [a], Astik Biswas [a,*], Anirban Bhowmick [b], Mahesh Chandra [b]

[a] Dept. of Electrical Engineering, National Institute of Technology, Rourkela, India
[b] Dept. of ECE, Birla Institute of Technology, Mesra, Ranchi, India

ARTICLE INFO

Article history:
Received 15 December 2013
Received in revised form
24 April 2014
Accepted 24 April 2014
Available online 28 May 2014

Keywords:
Speech recognition
Wavelet packets
ERB scale
WERBC
WMFCC
Phoneme recognition

ABSTRACT

In recent years wavelet transform has been found to be an effective tool for time–frequency analysis. Wavelet transform has been used as feature extraction in speech recognition applications and it has proved to be an effective technique for unvoiced phoneme classification. In this paper a new filter structure using admissible wavelet packet is analyzed for English phoneme recognition. These filters have the benefit of having frequency bands spacing similar to the auditory Equivalent Rectangular Bandwidth (ERB) scale. Central frequencies of ERB scale are equally distributed along the frequency response of human cochlea. A new sets of features are derived using wavelet packet transform's multi-resolution capabilities and found to be better than conventional features for unvoiced phoneme problems. Some of the noises from NOISEX-92 database has been used for preparing the artificial noisy database to test the robustness of wavelet based features.

1. Introduction

Speech as a medium of human to machine or machine to machine communication has been gaining popularity since the last few decades. Artificial intelligence cannot cultivate significantly without the improvement of automatic speech recognition. Most of the systems developed till now are based on the frequency domain analysis of the speech signal in a laboratory environment. However, speech recognition accuracy still degrades significantly in adverse real time situation and sensor mismatch conditions.

Automatic Speech Recognition (ASR) system comprises front end processing and back end processing. Front end encompasses various feature extraction and noise compensation techniques. Back end have different types of acoustic, language and pronunciation. Feature extraction is a technique of extracting optimum maximal information from a phoneme which gives maximum discrimination between phoneme classes. Feature extraction technique should be robust enough to perform well in different environmental conditions as well as sensor mismatch conditions. Apart from wavelet based feature extraction techniques some of the commonly used feature extraction techniques are Mel Frequency Cepstral Coefficients (MFCCs) [1], Linear Prediction based Cepstral Coefficients (LPCCs) [2], Gammatone Feature Cepstral Coeffficients (GFCC) [3,4], perceptual linear prediction [5]. Feature extraction techniques should be preceded by Fourier Transform (FT) in order to obtain its speech spectrum. Having a uniform resolution over the frequency plane windowed FT or the Short Time Fourier Transform (STFT) technique is not suitable to recognize some of the phonemes such as stops. It is difficult to detect a short event like burst in a slowly time varying signal by using STFT technique. To overcome this problem, wavelet packets (WPs) and local cosine transforms have helped in feature extraction [6–8].

Wavelet Packets (WPs) [9–11] are considered to have important signal representation schemes impacting compression, detection and classification [12,13]. WPs are extensively used in the analysis of pseudo-stationary time series processes and quasi-periodic random fields, such as the acoustic speech process [14,15]. WPs can be used effectively to describe a rich coverage of signal-space decomposition as well as providing a way for generating sub-band dependent partitions of the observation space. In conclusion, WPs induce a family of structural filter-banks with rich coverage of time–frequency characteristics that has the potential for enriching the way conventional MFCC features describe the short term behavior of the acoustic speech process.

WPs and multi-rate filter bank analysis have been adopted to improve the performance of conventional features by dividing the

* Corresponding author.
E-mail addresses: pksahu@nitrkl.ac.in (P.K. Sahu), astikbiswas@live.com (A. Biswas), anirban.bhowmick@outlook.com (A. Bhowmick), shrotriya69@rediffmail.com (M. Chandra).

Peer review under responsibility of Karabuk University.

http://dx.doi.org/10.1016/j.jestch.2014.04.004

frequency axis analogs to MEL scale frequency resolution in the context of ASR [7,16–19]. They used the Daubechies (db) two channel filter (TCF) which is reported to enhance the recognition performance for specific phone subcategories (stops and unvoiced speech) in a portion of the TIMIT. Choueiter and Glass (2007) [15] explored the problem of two-channel filter-bank design and they proposed the novel framework of rational filter-banks. Main focus of this work was to improve the frequency selectivity with respect to the conventionally adopted Daubechies WPs by designing a type of MEL frequency filter-bank structure. Improved performances were achieved in a simplified phone-segmented classification task with respect to MFCCs. Farooq et al. (2010) [17] used wavelet transform-based feature extraction technique by taking into account temporal as well as frequency band energy variations for Hindi phoneme recognition. This feature extraction technique performed better than MFCC features in a simplified phone classification. Litvin and Cohen (2011) [19] have shown that wavelet based bark scale aligned WP decomposition improves the performance of single-channel source separation of audio signals. Recently Pavez and Silva (2012) [18] have shown that wavelet based wavelet Packet Cepstral Coefficients (WPCC's) have shown concrete results that complement the previous work on supporting the use of WPs as a feature extraction techniques for ASR.

In this paper WP based features are wavelet based, in which the frequency axis is divided analogs to the Equivalent Rectangular Bandwidth (ERB) [20] scale frequency resolution. This ERB scale was originally designed to model human cochlear filtering [21]. ERB scale frequency resolution can be used to approximate center frequencies and the bandwidth of each Gammatone filter in GFCC. Frequency axis has been divided according to ERB scale to follow the response of human cochlea. In this paper it has been tried to take the advantage of auditory ERB filter-bank as well as WP to extract the coefficients at a certain frequency of interest. This technique attempts to reduce the articulation effect in the phoneme features. Recently we have shown the effectiveness of these ERB features for Hindi consonant recognition applications [22]. The performance of this feature technique have been tested with TIMIT database. Further, these features have proved more robust in presence of babble, volvo, factory and white noises. The performance of the wavelet based feature is compared with wavelet like MFCC(WMFCC) [8,17], MFCC and GFCC.

The rest of the article is organized as follows: Section 2 gives brief overview of wavelet based feature extraction technique. Section 3 provides brief overview of TIMIT database. Section 4 covers the details of experiments performed and result obtained for phone recognition task. Finally, the conclusions of the experiment are drawn in Section 5.

2. ERB like WP decomposition and feature extraction

Refs. [11,16] can be referred for detail description of wavelet analysis. The 24 sub-band wavelet packet tree is derived which approximate the ERB scale division as shown in Fig. 1. The

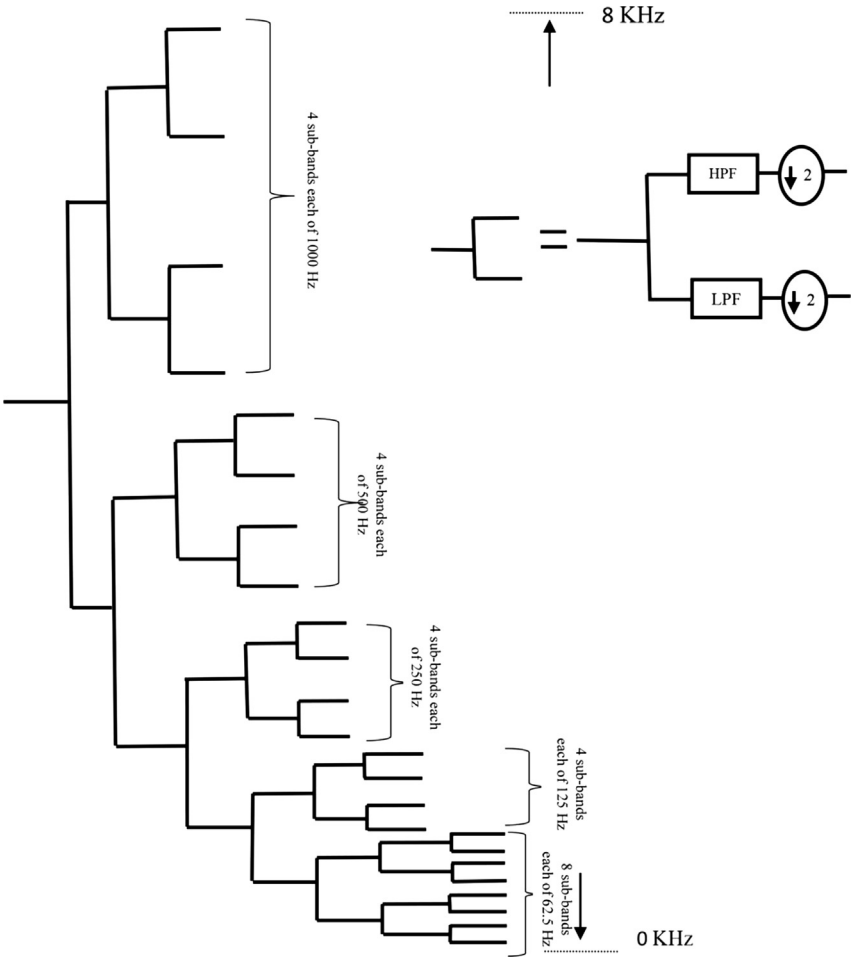


Fig. 1. 24 sub-band wavelet packet tree based on ERB scale.

mathematical relationship between the center frequency (f_c) and the ERB of an auditory filter is given by:

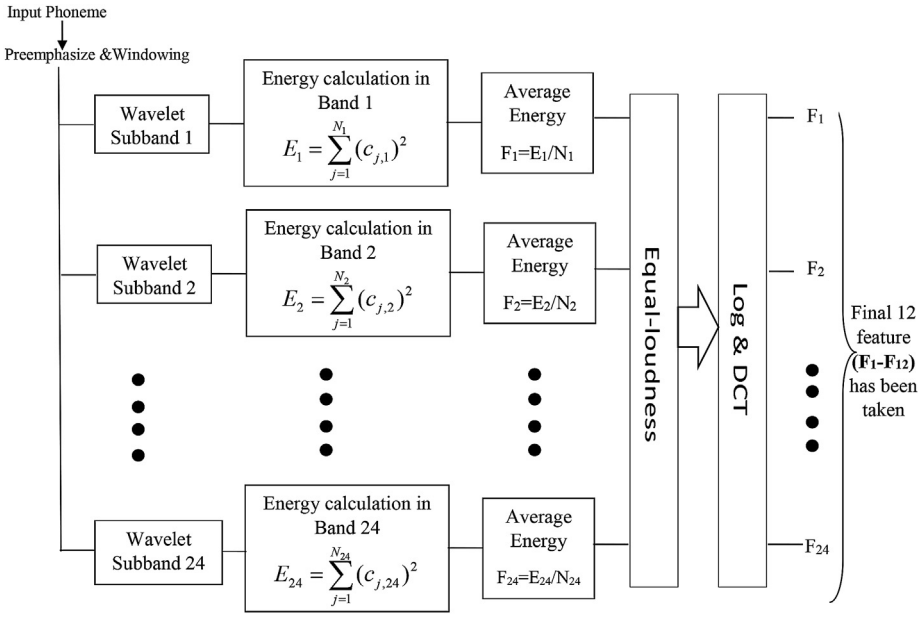
$$\mathrm{ERB} = 24.7\left(\frac{4.37 f_c}{1000} + 1\right) \quad (1)$$

The WP decomposition has been achieved by using a pair of conjugate mirror filters [7]. Thus decomposing signal into two frequency bands such as lower frequency band (approximation coefficients) and higher frequency band (detail coefficients). Low frequency band is used for further decomposition. Wavelet packet tree has been formed by cascading two channel filter bank into various levels.

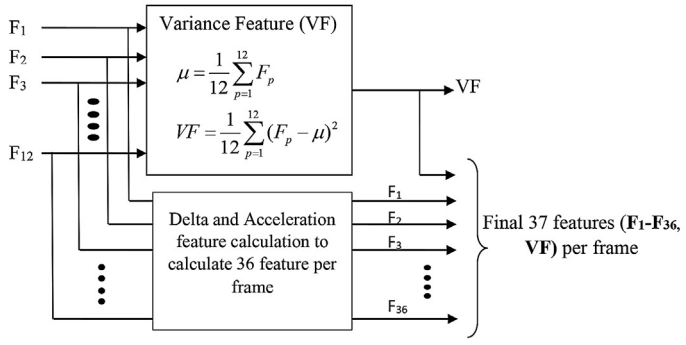
The speech in the TIMIT database is sampled at 16 kHz, giving an 8 kHz bandwidth signal. The ability of the admissible wavelet packet transform is used to divide a signal into ERB filter like 24-sub-bands. A frame size of 24 ms with 10 ms skip rate has been used to derive wavelet packet based ERB cepstral features (WERBC). Initially, hamming window is applied on each frame. Then, whole frequency band is decomposed using full 3-level wavelet packet decomposition to get eight sub-bands each of 1 kHz. Further one level WP decomposition is applied to lowest sub-band of 0–1 kHz to decompose the frequency band into two sub-bands each of 500 Hz. The frequency band of 0–500 Hz is further divided into eight sub-bands each of 62.5 Hz by using full 3 level WP decomposition. The resulting sub-band division finely emphasizes frequencies between 0 and 500 Hz which normally contains large portion of signal energy. Next, 500–1000 Hz, and 1–2 kHz frequency band have been decomposed using full 2 level WP decomposition to get sub-bands each of 125 Hz and 250 Hz. Then 2–3 kHz and 3–4 kHz is frequency band is decomposed using full 1-level WP decomposition to get sub-bands each of 500 Hz. Four frequency bands 4–5 kHz, 5–6 kHz, 6–7 kHz, & 7–8 kHz have been kept unchanged. Lastly, 24 total frequency sub-bands have been achieved. The center frequency obtained of each filter using WP decomposition is given in Table A1 (appendix). It can be noted from Table A1 that for first 20 sub-bands, wavelet frequency partitioning are similar alike the auditory ERB scale but the last 4 sub-bands differ from the ERB scale. However voice signals ranges upto 4000 Hz and most of the speech energy lies below 1500 Hz. Hence it is expected that these wavelet packet filters can extract certain information from speech signal by employing ERB like frequency decomposition. After performing the decomposition by WP of a phoneme, energy in each of the frequency bands has been calculated by:

$$\langle S_i \rangle_k = \frac{\sum \left[w_\psi(x,k)_i \right]^2}{N_i} \quad (2)$$

where, $W_\psi(x)$ is the WP transform of signal x, i is the sub-band frequency index ($1 \le i \le M$), k represents the temporal frame and



(a) Block diagram of wavelet based feature extraction technique



(b) Calculation of variance feature along with EDA feature

Fig. 2. ERB like wavelet based feature extraction process.

N_i is the number of coefficient in ith sub-band. The log of equal loudness weighted energy has been applied resulting into 24 coefficients. Finally discrete cosine transform (DCT) has been applied on these 24 coefficients to de-correlate the filter bank energies and first 12 coefficients have been taken as features.

To capture the dynamic information of speech signal, static feature vector has been added with delta and acceleration coefficients. In this way total 36 EDA (energy, delta & acceleration) features per frame were obtained. To make wavelet based feature more robust in noisy environment another additional feature is also calculated based on the variance of the energy features as shown in Fig. 2. Prior to computation of variance feature (VF), average sub-band energy (μ) has been calculated. VF helps in the recognition of phonemes, as the variance is not altered by a constant addition, which may occur due to noise. Thus finally a total of 37 features are obtained per frame.

3. TIMIT database

The TIMIT corpus was adopted for all the experiments presented in this work. TIMIT is one of the standard corpus used to evaluate the performance of new techniques in ASR because it is a phonetically balanced database and has good coverage of speakers and dialects. All of these make TIMIT a sufficiently challenging corpus to evaluate new ASR methods, which justifies its wide adoption by the community. The TIMIT corpus consists of 6300 utterances for 8 major dialects of the United States. There are 630 different speakers, each one speaking 10 sentences. For this experiment Dialect region DR1, DR2, DR3 and DR4 from training set were chosen to train the system. Recognition of nasals (|*m*|, |*n*| & |*ng*|), unvoiced fricatives (|*f*|, |*sh*|, |*s*| & |*th*|), voiced fricatives (|*z*|, |*v*|, |*zh*| & |*dh*|), liquids (|*l*|, |*r*|, |*y*| & |*w*|), unvoiced stops (|*p*|, |*t*|& |*k*|), and voiced stops (|*b*|, |*d*| & |*g*|) have been carried out. The dialect region DR1, DR2, DR3 and DR4 from complete test set were used for testing. The individual phoneme composition used in this experiment to study the performance of wavelet based feature set is provided in Appendix (Table A2 and A3). The speech signal was pre-emphasized to ensure that all formants of acoustic signals have similar amplitudes so that they get equal importance in subsequent processing stages.

4. Experimental setup and results

Here similar kind of experimental framework as adopted by Messaoud and Hamida (2010) [23] is followed. One model was created for every phones and each HMM model has three emitting states with eight Gaussian mixtures. TIMIT phone-level annotation was used to initialize HMM parameters, followed by Viterbi alignment to improve the state-time correspondence. The Baum–Welch algorithm was then applied at the sentence level. Then, triphone context dependent HMMs [24] were made using a phonetic decision class tree. Decoding is performed by compiling a network of all vocabulary phonemes in parallel within a loop [25]. Once compiled, the whole recognition network can be used in a conventional Viterbi decoder to classify the phoneme into their respective classes for an unknown input utterance. Phoneme Recognition Accuracy (PRA) is calculated by the following equation:

$$\mathrm{PRA}\ (\%) = 100\ (\%) - \mathrm{PER}\ (\%) \tag{3}$$

where Phoneme Error rate (PER) is given by:

$$\mathrm{PER}(\%) = \frac{(\mathrm{Substitutions} + \mathrm{Deletions} + \mathrm{Insertions})}{\mathrm{Total\ \ Phoneme}} \times 100 \tag{4}$$

Table 1
Phoneme recognition accuracy for different system with baseline features.

Phoneme	System					
	CI		CD		Gain	
	WMFCC	WERBC	WMFCC	WERBC	WMFCC	WERBC
Nasals	70.45	71.73	76.20	78.09	5.75	6.36
Voiced stop	71.88	71.12	79.27	79.55	7.39	8.43
Unvoiced stop	76.70	78.85	83.60	85.08	6.90	6.23
Voiced fricative	73.30	74.22	81.80	82.90	8.50	8.68
Unvoiced fricative	81.20	82.95	87.13	89.25	5.93	6.30
Liquids	72.64	74.78	78.07	81.15	5.43	6.37
Avg	**74.36**	**75.61**	**81.01**	**82.67**	**6.65**	**7.06**

4.1. Baseline recognition result

Baseline recognition tests have been carried out by using the conventional 36 MFCC and GFCC features. MFCC and GFCC features have been derived using a frame size of 24 ms with 10 ms skip rate. Initially, the experiment was started with context-independent (CI) phoneme model then switched to context-dependent (CD) phoneme recognition experiment. The results obtained from this CI and CD experiment are shown in Table 1. Results reveal that CD phoneme models have shown significant improvement over CI model. This mismatch of results when switching from CI to CD system could be explained by the fact that the articulatory information had taken a big advantage of modeling coarticulation phenomena in CD system to improve results. Phones are highly affected by the neighboring phonetic contexts and CD model has taken care of these facts. Result shows that average PRA (%) with GFCC is better than MFCC features, because it takes the advantage of gammatone filter bank which was designed according to the model of human cochlear filtering. In the next sub-section, the results with WMFCC and wavelet based features will be presented.

4.2. Performance evaluation of wavelet based feature

These experiments have been carried out to compare the performance of our recognition system using the new set of wavelet features with conventional MFCC, GFCC and WMFCC. A frame size of 24 ms with 10 ms skip rate has been used to derive wavelet packet based feature. WMFCC and WERBC features have been derived using db24 mother wavelet. Table 2 shows the performance of wavelet packet based features. It is observed that the recognition performance of WP derived features is better than MFCC and GFCC features except voiced fricative phoneme class. MFCC and GFCC features are superior because it uses the STFT, having sine and cosine basis, which are more efficient to extract the periodic structure from a signal. WP derived features performed better for the stop classes because stops have a sudden burst of high frequency that cannot be detected perfectly due constant

Table 2
Phones recognition accuracy for different system with wavelet packet based features.

Phoneme	System					
	CI		CD		Gain	
	WMFCC	WERBC	WMFCC	WERBC	WMFCC	WERBC
Nasals	70.45	71.73	76.20	78.09	5.75	6.36
Voiced stop	71.88	71.12	79.27	79.55	7.39	8.43
Unvoiced stop	76.70	78.85	83.60	85.08	6.90	6.23
Voiced fricative	73.30	74.22	81.80	82.90	8.50	8.68
Unvoiced fricative	81.20	82.95	87.13	89.25	5.93	6.30
Liquids	72.64	74.78	78.07	81.15	5.43	6.37
Avg	**74.36**	**75.61**	**81.01**	**82.67**	**6.65**	**7.06**

Table 3
Percentage recognition gain achieved with WEBRC compared to other features.

Phoneme	System					
	CI			CD		
	MFCC	GFCC	WMFCC	MFCC	GFCC	WMFCC
Nasals	4.03	2.38	1.28	3.41	1.19	1.89
Voiced stop	1.92	−1.74	−0.76	3.10	0.99	0.28
Unvoiced stop	4.20	5.80	2.15	4.20	5.24	1.48
Voiced fricative	−0.86	−1.73	0.92	1.64	−0.63	1.10
Unvoiced fricative	4.70	5.37	1.75	4.12	3.40	2.12
Liquids	4.93	2.90	2.14	5.88	3.20	3.08
Avg	**3.15**	**2.16**	**1.25**	**3.73**	**2.23**	**1.66**

resolution in the time–frequency plane of STFT. These features can be easily captured by wavelet analysis due to its multi-resolution property. Further, wavelet based feature extraction technique proved to be superior over other feature extraction technique in most of the phoneme classes because it takes the advantage of wavelet analysis along with auditory ERB scale. Table 3 shows the PRA gain achieved with WERBC compared to other features. Satisfactory improvement was achieved with wavelet based features because wavelet packet decomposition was carried out according to the ERB scale which seeks to segregate target speech from a composite auditory scene. The detailed phoneme error analysis is presented in Table 4. The substitution error was detailed in two errors: inter and intra substitution phones groups.

It can be seen from Table 4 that the inter substitution rate (misclassified to other phoneme category) of wavelet features for unvoiced stop classes is significantly lower compared to STFT based features. It shows the efficiency of features regarding the classification of unvoiced phonemes. Besides, in case of liquids and nasal group, enhancement shown in PER rate is especially attributed to inter substitution error rate which shows a significant fall compared to the baseline system.

4.3. Performance evaluation in dialect mismatch condition

To study the effectiveness of wavelet based feature in dialect mismatch condition complete test set from dialect DR5, DR6 & DR7 were taken. Table 5 has shown the recognition performance of all features in dialect mismatch condition. Table 5 shows the robustness of ERB based features in the dialect mismatch condition. The performance of MFCC features drops down relatively by 3.6% due to the fact that Mel scale might be less superior to track the dialectal changes which slows down phoneme recognition. By use of time frequency analysis property of WP, WMFCC shows some improvement in recognition performance. However, auditory GFCC features have shown better recognition efficiency which proves the adaptability of ERB to the dialect mismatch condition. WERBC found to be best in the dialect mismatch condition by taking the advantage of auditory ERB like sub-band wavelet packet decomposition.

Table 5
Performance evaluation of wavelet based feature in dialect mismatch condition. Relative change (%) is shown in parentheses compared to performance of clean training condition.

Phoneme	MFCC	GFCC	WMFCC	WERBC
Nasals	71.98 (−3.62)	74.89 (−1.34)	75.32 (−1.16)	77.39 (−0.90)
Voiced stop	74.03 (−3.67)	77.28 (−1.63)	77.68 (−2.01)	78.85 (−0.88)
Unvoiced stop	76.95 (−5.10)	77.74 (−2.64)	80.86 (−3.28)	83.18 (−2.24)
Voiced fricative	81.76 (−1.81)	83.85 (−0.72)	80.18 (−1.99)	81.39 (−1.83)
Unvoiced fricative	82.43 (−3.37)	84.09 (−2.06)	85.33 (−2.07)	88.28 (−1.09)
Liquids	73.20 (−3.52)	76.22 (−1.34)	76.56 (−1.94)	80.33 (−1.02)
Avg.	**76.72 (−3.5)**	**79.02 (−1.62)**	**79.32 (−2.09)**	**81.57 (−1.34)**

4.4. Performance evaluation in noisy environment

Finally noisy phoneme recognition task has been carried out to evaluate robustness of wavelet based features. To evaluate the robustness of the wavelet based features, the babble, factory, volvo and white noises from NOISEX-92 database were added to clean signal at different SNR levels. Context Dependent (CD) Phoneme recognition accuracy was evaluated for SNRs in the range from 0 dB to 20 dB. The average PRA of English phoneme recognizer under different level and type of noise is shown in Fig. 3. This clearly shows the improved performance of the WP derived features for English phonemes over MFCC and GFCC features, especially for low SNR values. WP derived features are less sensitive to noise and it can extract the coefficients at a certain frequency of interest. Further result shows that ERB filter WP derived features are superior compared to WMFCC features. Wavelet based feature take the advantage of wavelet analysis as well as it is designed according to the frequency response of human cochlea (ERB scale). Due to differences in the stationary characteristics of speech and noisy signals, the ERB filter bank is less sensitive to noise and concentrate on speech signal.

4.5. Performance evaluation on overall TIMIT phone set

To study the overall performance of wavelet based features all 39 TIMIT phone set has been used. Context dependent triphone models have been prepared to evaluate the performance of wavelet based features. Table 6 shows the average phone recognition accuracy of different front end features in clean as well as noisy conditions. Average performance of all four types of noises has been reported in Table 6. It is interesting to see that the performance of STFT based features have significantly improved with overall clean TIMIT phone set. STFT based technique shown better performance compared to consonant recognition problem because of the inclusion of vowels and other voiced phonemes. It is well known fact that STFT based techniques are more superior to extract the periodic information from voiced phonemes. But inclusion of noise has a high impact on STFT based features especially on MFCC. While GFCC is purely auditory based method and center frequencies are distributed according to the ERB scale, which can focus and separate target speech in composite auditory scene. This proves the effectiveness of ERB

Table 4
Detailed PER (%) analysis with different types of error.

Phoneme	Analysis															
	Deletion (%)				Intra substitution (%)				Inter substitution (%)				Insertion (%)			
	MFCC	GFCC	WMFCC	WERBC	MFCC	GFCC	WMFCC	WERBC	MFCC	GFCC	WMFCC	WERBC	MFCC	GFCC	WMFCC	WERBC
Nasals	5.20	4.85	5.57	5.05	6.35	8.88	10.47	9.78	8.88	6.52	4.65	3.65	4.89	2.85	3.11	3.43
Voiced Stop	5.92	3.26	4.75	4.40	10.76	9.98	7.30	9.35	2.85	2.95	4.14	2.20	4.02	5.25	4.54	4.50
Unvoiced Stop	4.75	4.38	3.26	3.05	2.65	2.24	5.87	5.45	6.71	7.57	2.65	1.95	5.01	5.97	4.62	4.47
Voiced Fricative	4.27	3.78	3.90	2.79	8.35	6.05	5.25	5.90	1.58	1.78	4.27	4.75	4.54	4.86	4.78	3.66
Unvoiced Fricative	3.58	2.47	3.12	1.95	2.65	3.28	4.21	4.48	5.05	5.90	2.15	1.85	3.59	2.50	3.39	2.47
Liquids	5.85	4.95	4.54	3.24	6.28	5.70	11.24	9.98	7.57	6.28	0.79	1.43	5.03	5.12	5.36	4.20

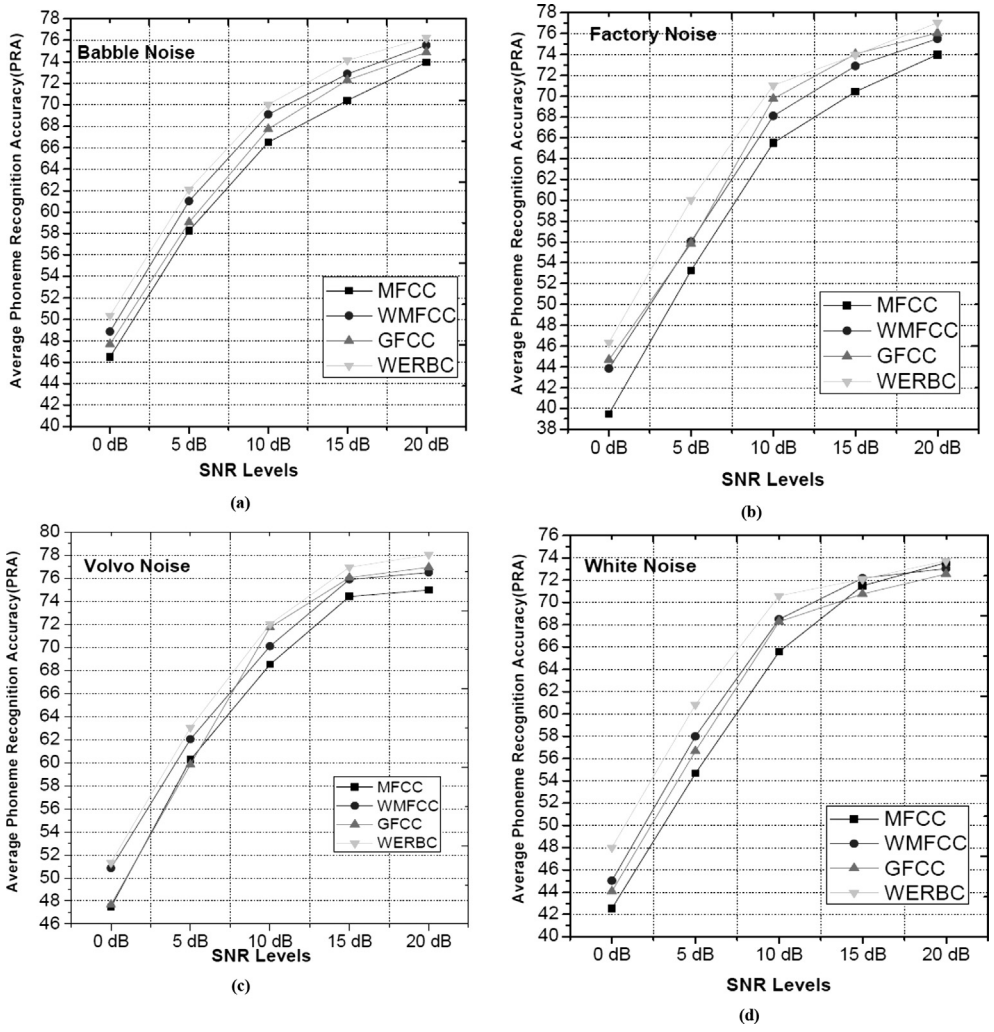


Fig. 3. Average PRA(%) in presence of different types of noises (a) babble noise, (b) factory noise, (c) volvo noise, (d) white noise.

scale while working in noisy condition. WERBC features have been found better especially in low SNR conditions. WERBC features outperformed other features by ample margin in noisy condition. This proves the effectiveness of WERBC in environmental mismatch condition between train and test data. As described in Section 2, it has been tried to increase the frequency resolution in the low-frequency range. This is well known fact that the discriminative information of the speech signal is embedded in lower frequency bands. The speech production-perception hypothesis suggests that for an optimal communication design, maximum signal energy should be embedded in lower frequency region where more perception (frequency discrimination) is available.

5. Conclusion

A new set of auditory ERB like wavelet features has been presented by keeping same number of sub-bands as that of ERB filter. Experiments have been carried out in sequential steps to see the performance of new wavelet based features. Comparative study with baseline systems is also presented to show the robustness of the wavelet based features. The multi-resolution property of wavelet allows for a better modeling of phoneme classes, especially for voiceless class. The performance of the new feature is studied for the task of phoneme recognition. Wavelet based features have shown an overall improvement in recognition performance for English phoneme as compared to WMFCC and STFT based features. WERBC is found to be superior compared to the WMFCC especially in case of noisy condition. The speaker independent results show considerable improvement in recognition of the phoneme classes tested with TIMIT database. Further, the wavelet based features are found to be robust in presence of different noises.

Table 6
Average performance recognition accuracy on overall TIMIT phone set.

	Clean	0 dB	5 dB	10 dB	15 dB	20 dB
MFCC	80.26	42.86	54.86	66.28	74.26	76.89
GFCC	80.84	46.41	58.21	68.48	74.88	77.54
WMFCC	79.95	47.05	57.24	68.71	72.89	75.32
WERBC	81.78	48.76	61.08	73.32	73.08	76.81

Acknowledgments

We are thankful to the respected reviewers and honorable editor for providing important suggestions and constructive comments which have helped us in improving the quality of the paper.

Appendix

Table A1
Comparison of center frequencies (Hz) of 24 uniformly spaced ERB scale and wavelet packet sub-band.

Filters	ERB scale	Wavelet sub-band	Filters	ERB scale	Wavelet sub-band	Filters	ERB scale	Wavelet sub-band
1	50	62.5	9	632.83	625	17	2433.98	2500
2	92.23	125	10	763.35	750	18	2837.29	3000
3	140.86	187.5	11	913.62	875	19	3301.7	3500
4	196.85	250	12	1086.66	1000	20	3836.44	4000
5	261.33	312.5	13	1285.92	1250	21	4452.17	5000
6	335.57	375	14	1515.35	1500	22	5161.17	6000
7	421.06	437.5	15	1779.52	1750	23	5977.56	7000
8	519.49	500	16	2083.71	2000	24	6917.58	8000

Table A2
Composition of phonemes (number of tokens) used in the experiment to train the system using 4 dialect region.

	Unvoiced stop	Voiced stop	Unvoiced fricative	Voiced fricative	Nasals	Liquids
DR1	926	653	1044	722	1000	1389
DR2	1984	1261	2129	1376	2062	2886
DR3	1866	1286	2058	1465	1979	2864
DR4	1776	1079	1831	1335	1827	2470
Total	6552	4279	7062	4898	6868	9609

Table A3
Composition of phonemes (number of tokens) used in the experiment to test the system using 7 dialect region of TIMIT.

	Unvoiced stop	Voiced stop	Unvoiced fricative	Voiced fricative	Nasals	Liquids
Dialect specific testing						
DR1	244	193	326	202	298	419
DR2	678	459	710	450	701	1027
DR3	617	431	695	497	683	1048
DR4	531	378	578	451	560	927
Total	2070	1461	2309	1600	2242	3421
Dialect mismatch condition						
DR5	941	650	998	738	965	1486
DR6	256	175	330	191	298	429
DR7	597	412	636	432	631	944
Total	1794	1237	1964	1361	1894	2859

References

[1] S.B. Davis, P. Mermelstein, Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences, IEEE Trans. Acoust. Speech Signal Process. (1980) 357–366. ASSP-28.

[2] E. Wong, S. Sridharan, Comparison of linear prediction cepstrum coefficients and mel-frequency cepstrum coefficients for language identification, Proc. IEEE Int. Symp. Intell. Multimed. Video Speech Process. (2001) 95–98.

[3] A. Biswas, P.K. Sahu, A. Bhowmick, M. Chandra, Hindi vowel classification using GFCC and formant analysis in sensor mismatch condition, WSEAS Trans. Syst. 13 (2014) 130–143.

[4] Y. Shao, S. Srinivasan, Z. Jin, D. Wang, A computational auditory scene analysis system for speech segregation and robust speech recognition, Comput. Speech Lang. Elsevier 24 (2010) 77–93.

[5] H. Hermansky, Perceptual linear predictive (PLP) analysis, J. Acoust. Soc. Am. 87 (1990) 1738–1752.

[6] C. Long, S. Datta, Wavelet Based Feature Extraction for Phoneme Recognition, 4th Int. Conf. Spok. Lang. Process., Philadelphia (USA), 1996, pp. 264–267.

[7] O. Farooq, S. Datta, Mel filter-like admissible wavelet packet structure for speech recognition, IEEE Signal Process. Lett. 8 (2001) 196–198.

[8] R. Sarikiya, B. Pellom, J.H.L. Hansen, Wavelet Packet Transform Features with Application to Speaker Identification, Proc. IEEE Nord. Signal Process. Symp., 1998, pp. 81–84

[9] M. Vetterli, J. Kovacevic, Wavelet and Subband Coding, Prentice Hall, Englewood Cliffs, New Jersey, 1995.

[10] S.A. Mallat, Wavelet Tour of Signal Processing, 3rd ed., Academic Press, 2009.

[11] S. Mallat, A theory for multiresolution signal decomposition: the wavelet representation, IEEE Trans. Pattern Anal. Mach. Intell. 11 (1989) 674–693.

[12] C. Scott, R. Nowak, Templar: a wavelet-based framework for pattern learning and analysis, IEEE Trans. Signal Process. 52 (2004) 2264–2274.

[13] K. Etemad, R. Chellapa, Separability-based multiscale basis selection and feature extraction for signal and image classification, IEEE Trans. Image Process. 7 (1998) 1453–1465.

[14] J. Silva, S. Narayanan, Discriminative wavelet packet filter bank selection for pattern recognition, IEEE Trans. Signal Process. 57 (2009) 1796–1810.

[15] G. Choueiter, J. Glass, An implementation of rational wavelets and filter design for phonetic classification, IEEE Trans. Audio Speech Lang. Process. 15 (2007) 939–948.

[16] O. Farooq, S. Datta, Wavelet based robust sub-band features for phoneme recognition, IEE Proc. Vision, Image Signal Process. 151 (2004) 187–193.

[17] O. Farooq, S. Datta, M.C. Shrotriya, Wavelet sub-band based temporal features for robust Hindi phoneme recognition, Int. J. Wavelets Multiresolut. Inf Process. 8 (2010) 847–859.

[18] E. Pavez, J.F. Silva, Analysis and design of wavelet-packet cepstral coefficients for automatic speech recognition, Speech Commun. Elsevier 54 (2012) 814–835.

[19] Y. Litivin, I. Cohen, Single-channel source separation of audio signals using bark scale wavelet packet decomposition, J. Signal Process. Syst. Springer 65 (2011) 339–350.

[20] B.C.J. Moore, An Introduction to the Psychology of Hearing, Academic Press, San Diego, 2003.

[21] R.D. Patterson, I. Nimmo-Smith, J. Holdsworth, P. Rice, An Efficient Auditory Filterbank Based on the Gammatone Function, Appl. Psychol. Unit, Cambridge University, 1988.

[22] A. Biswas, P.K. Sahu, M. Chandra, Admissible Wavelet Packet Features based on Human Inner Ear Frequency Response for Hindi Consonant Recognition, Comput. Electr. Eng.,Elsevier 40 (2014) pp. 1111–1122, 2014.

[23] Z.B. Messaoud, A.B. Hamida, Combining formant frequency based on variable order LPC coding with acoustic features for TIMIT phone recognition, Int. J. Speech Technol. Springer 14 (2010) 393–403.

[24] K.F. Lee, H.W. Hon, Speaker-independent phone recognition using hidden Markov models, IEEE Trans. Acoust. Speech Signal Process. 37 (1989) 1641–1648.

[25] N. Kumar, A.G. Andreou, A Generalization of Linear Discriminant Analysis in Maximum Likelihood Framework, Johns Hopkins University, 1996.