

Feature extraction technique using ERB like wavelet sub-band periodic and aperiodic decomposition for TIMIT phoneme recognition

Astik Biswas · P. K. Sahu · Anirban Bhowmick · Mahesh Chandra

Received: 20 December 2013 / Accepted: 19 May 2014 / Published online: 4 June 2014
© Springer Science+Business Media New York 2014

Abstract In the recent years, wavelet transform has been found to be an effective tool for the time–frequency analysis for non-stationary and quasi-stationary signals such as speech signals. In the recent past, wavelet transform has been used as feature extraction in speech recognition applications. Here we propose a wavelet based feature extraction technique that signifies both the periodic and aperiodic information along with sub-band instantaneous frequency of speech signal for robust speech recognition in noisy environment. This technique is based on parallel distributed processing technique inspired by the human speech perception process. This frontend feature processing technique employs equivalent rectangular bandwidth (ERB) filter like wavelet speech feature extraction method called Wavelet ERB Sub-band based Periodicity and Aperiodicity Decomposition (WERB-SPADE), and examines its validity for TIMIT phone recognition task in noisy environments. The speech sound is filtered by 24 band ERB like wavelet filter banks, and then the equal loudness pre-emphasized output of each band is processed through comb filter. Each comb filter is designed individually for each frequency sub-band to decompose the signal into periodic and aperiodic features. Thus it takes

the advantage of the robustness shown by periodic features without losing certain important information like formant transition incorporated in aperiodic features. Speech recognition experiments with a standard HMM recognizer under both clean-training and multi-training condition training is conducted. Proposed technique shows more robustness compared to other features especially in noisy condition.

Keywords ERB · WERB-SPADE · Wavelet sub-band · Instantaneous frequency · Phoneme recognition

1 Introduction

In the recent past, the use of speech as a possible interface with machines/computer has become popular. Significant research has been carried out to improve the performance of robust speech recognition system in the past couple of decades. However, most of these systems developed by both academicians and industry are based on the short time Fourier transform (STFT) for the analysis of speech signal. These systems have shown adequate recognition performance with clean data, keeping same acoustic conditions. Nevertheless, speech recognition accuracy still degrades significantly in noisy environments and sensor mismatch conditions. This fact has motivated to develop a new feature extraction methods to represent more robust features.

Mel frequency cepstral coefficients (MFCCs) (Davis and Mermelstein 1980) is the most widely used front-end feature extraction technique, which represents the spectral shapes of input signals and it is very sensitive to noise. However, human being can follow and segregate the target speech in complex acoustic environment until and unless noise becomes too high. This fact have inspired researchers to develop more robust speech feature extraction method according to

A. Biswas (✉) · P. K. Sahu
Department of Electrical Engineering, National Institute of Technology, Rourkela, India
e-mail: astikbiswas@live.com

P. K. Sahu
e-mail: pksahu@nitrkl.ac.in

A. Bhowmick · M. Chandra
Department of Electronics & Communications, Birla Institute of Technology, Mesra, Ranchi, India
e-mail: anirban.bhowmick@outlook.com

M. Chandra
e-mail: shrotriya69@rediffmail.com

the human auditory system, such as perceptually linear prediction analysis (PLP) (Hermansky 1990), Gammatone frequency cepstral coefficient (GFCC) (Shao et al. 2010) and other auditory filter based methods (Gao et al. 1992; Li et al. 2001). These results suggest that the outputs of the auditory filters provide a better representation of speech in complex acoustic environment and sensor mismatch condition. However all the feature extraction techniques mentioned earlier is STFT based method and it is a well-known fact that the windowed FT or the STFT having uniform resolution over the time frequency plane. Due to this reason, it is difficult task to classify unvoiced phonemes especially ‘stops’. It is not an easy task to detect a short event like burst in a slowly time varying signal especially in the presence of noise. To overcome this problem, WPs and local cosine transforms have been proposed for the feature extraction technique (Farooq et al. 2010; Farooq and Datta 2001; Long and Datta 1996). WPs are considered to have important signal representation schemes impacting compression, detection and classification. This characteristic of WP is particularly exciting for the study of the slowly time varying quasi-periodic signal like acoustic speech signal. The previously reported results motivate us to develop a new wavelet based feature extraction technique. Recently Biswas et al. 2014 have proposed equivalent rectangular bandwidth (ERB) like WP based cepstral feature (WERBC) for Hindi phoneme recognition. They have shown significant improvement with WERBC in Hindi consonant class. Wavelet packets (WP) (Farooq and Datta 2004; Mallat 1989) decomposition has been carried out according to the ERB scale frequency resolution. This ERB like Wavelet feature extraction. This ERB scale was originally designed to model human cochlear filtering (Patterson et al. 1988). The center frequency and the bandwidth of each gammatone filter in GFCC can be derived from the filter’s ERB. They have tried to follow the response of human cochlea by partitioning the frequency axis analogous to the ERB scale. They have taken the advantage of auditory ERB filterbank as well as WP can extract the coefficients at a certain frequency of interest. Due to the dynamic nature of the WP, exact bandwidth alike to the ERB scale cannot be obtained. The advantage of using WP is that it can divide the frequency axis and has uniform translation in time. Although they have found sufficient improvement in unvoiced consonant class but WERBC is failed to keep same impression in case of voiced phoneme class compared to STFT based class. Here we have adopted the WERBC proposed by Biswas et al. (2014) and modify it to improve the performance for voiced phoneme class.

The main focus of our research motivated by the periodicity and aperiodicity of sound. Several methods have been proposed (Ali et al. 2002; Kim et al. 1999) based on the response of the auditory nerve towards the periodic signals. Kajita and Itakura (1995) proposed cross-correlation based sub-band analysis, which focuses on periodicity of band passed speech

signal. Reported results have shown periodic property of the signal helps to improve the performance in noisy environment. However, speech signals consist not only strict to periodic signals like voiced parts of vowels and voiced consonants, but also of aperiodic signals like intrinsic variations in vowels and unvoiced phonemes such as fricatives and stops. Thus considering only periodicity of signal not sufficient, when it comes to study with aperiodic sounds. Motivated by this fact, Ishizuka and Miyazaki (2004) proposed a speech feature extraction method known as “Sub-band based Periodicity and Aperiodicity Decomposition (SPADE)”. They have proposed a scheme to divide input signals into sub-band signals using gammatone filterbank, and then decomposes into periodic and aperiodic feature set in the time domain. The development of this technique has been inspired by the auditory comb filter assumption (Cheveigne’ et al. 1997) and the study reported on auditory nerve characteristics in the time domain by Greenberg et al. (2004). The auditory comb filter assumption suggests that the human auditory system may perceive both the harmonic (periodicity), and the residue after canceling the harmonic part from acoustic signal (aperiodicity), which deviates from the dominant periodicity. The decomposition of speech signals into periodic and aperiodic part helps to make more robust feature without losing certain essential aperiodic information presented in speech signal. They have evaluated the performance of SPADE with AURORA-2J database in the presence of noise and claimed that proposed feature have outperformed MFCC. Later Ishizuka and Nakatani (2006) have expanded the SPADE analysis in frequency domain and have proposed new feature extraction technique named SPADE frequency domain Enhancement (SPADE-QUEEN). They combined their proposed front end technique with different noise compensation technique such as, spectral subtraction (Berouti et al. 1979; El-Fattah et al. 2013) or Wiener filtering (Adami et al. 2002; El-Fattah et al. 2013) and studied the performance of robust front end technique with AURORA 2J database. Reported result have confirmed the robustness of SPADE combined with different noise compensation technique.

Motivated by the performance of SPADE analysis, here we propose a new feature extraction technique named Wavelet ERB Sub-band based Periodicity and Aperiodicity Decomposition (WERB-SPADE). All the above mentioned analysis of SPADE based on the STFT based approach, and have evaluated on AURORA-2J continuous digit speech database. The performance of SPADE on different phonemes have not been reported. Thus here noisy phoneme recognition especially consonants are carried out. Further as mentioned earlier STFT is not suitable for unvoiced phonemes like stops because detecting sudden “bursts” in slowly time varying signal is really near to impossible task for STFT based technique. The situation gets more worsen for the STFT based feature extraction technique in noisy environ-

ment. Thus wavelet based feature extraction technique is adopted here to capture time-frequency localized information about the sudden changes in the speech waveform. This paper proposes a wavelet based feature extraction technique that decompose speech signals into periodic and aperiodic features for each sub band using ERB like wavelet filter banks and comb filters. To make the proposed feature more robust average sub-band periodic energy parameter is incorporated to the proposed feature vector. Further to capture the phase information of speech signal sub-band instantaneous frequency (IF) is calculated. Recent work on auditory models proves that certain nonlinear effects of active human inner ear cochlear processing might be simulated using the auditory frequency sub-bands IF (Hohmann and Kollmeier 2006). Sub-band IF has already been proven in noisy speech recognition problem (Stark and Paliwal 2008; Yin et al. 2011). Thus we incorporate additional subband IF features, derived from output of auditory ERB like wavelet filterbank. Phonetically balanced standard TIMIT (Garofolo et al. 1993) database is used to study the performance of proposed feature extraction technique.

The rest of paper is organized as follows: Section 2 describes the feature extraction procedure using ERB like admissible wavelet packet (AWP) decomposition. Section 3 describes the detailed WERB-SPADE feature extraction mechanism. Section 4 describes the experimental framework

adopted for this experiment. The performance of proposed technique on TIMIT phoneme recognition task is reported in Sect. 6. Section 8 draws the conclusion.

2 ERB like WP decomposition and feature extraction

The detailed description of wavelet analysis is beyond the focus of this paper. Hence, we request interested readers to refer articles like (Farooq and Datta 2004; Mallat 1989). The 24 sub-band wavelet packet tree is derived which approximate the ERB scale division as shown in Fig. 1 (Biswas et al. 2014). The WP decomposition achieved by using a pair of conjugate mirror filters (Farooq and Datta 2001). 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. By cascading the two channel filter bank into various levels the wavelet packet tree has been formed.

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 16 ms with 10 ms skip rate is used to derive wavelet packet based ERB cepstral features (WERBC). Initially, hamming win-

Fig. 1 24 subband wavelet packet tree based on ERB scale

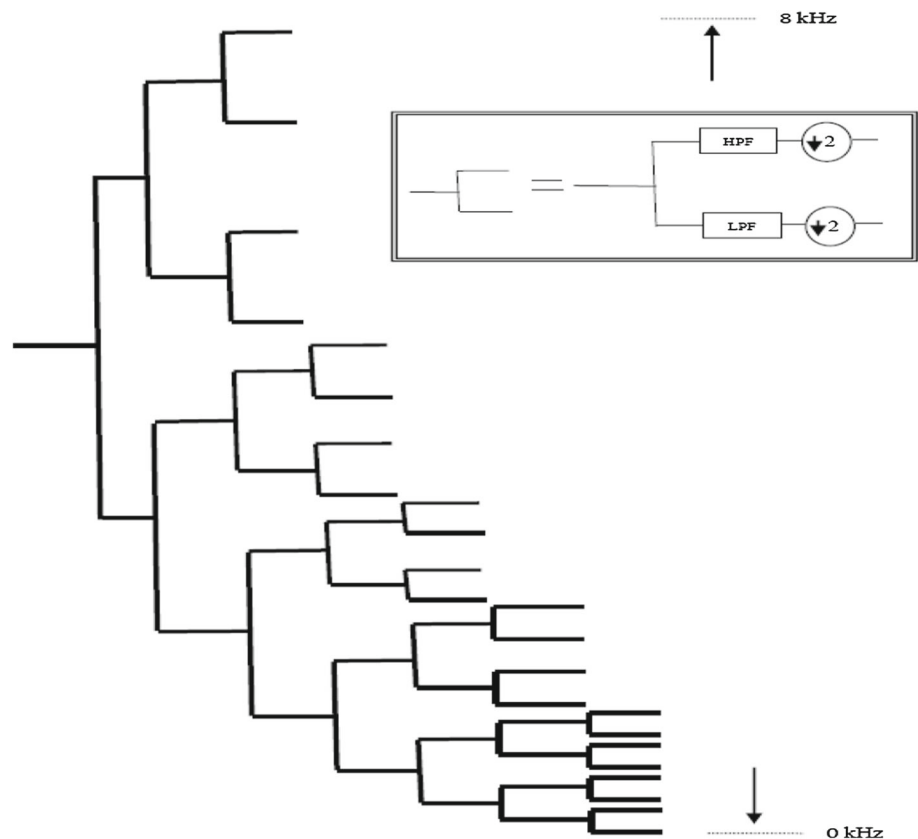


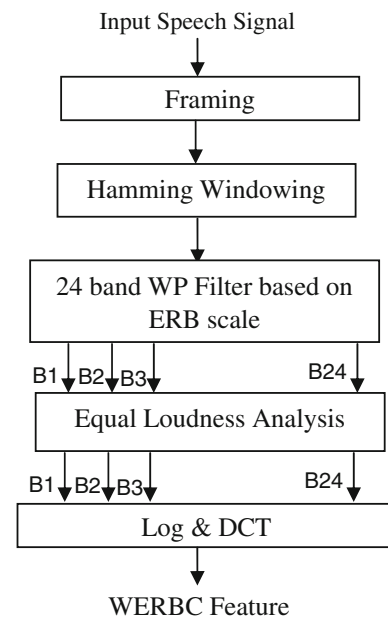
Table 1 Comparison of center frequencies (Hz) of 24 uniformly spaced ERB scale and wavelet sub-band features

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	2,433.98	2500
2	92.23	125	10	763.35	750	18	2,837.29	3000
3	140.86	187.5	11	913.62	875	19	3,301.70	3500
4	196.85	250	12	1,086.66	1000	20	3,836.44	4000
5	261.33	312.5	13	1,285.92	1250	21	4,452.17	5000
6	335.57	375	14	1,515.35	1500	22	5,161.17	6000
7	421.06	437.5	15	1,779.52	1750	23	5,977.56	7000
8	519.49	500	16	2,083.71	2000	24	6,917.58	8000

down 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 0–500 Hz which normally contains large portion of signal energy. Next, 500–1,000 Hz, and 1–2 kHz frequency band is decomposed using full 2 level WP decomposition to get sub-bands each of 125 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 is kept unchanged. Lastly, 24 total frequency sub-bands are achieved. The center frequency obtained of each filter using WP decomposition is given in Table 1. From the table it can be noted that for the first 20 sub-band wavelet frequency partitioning are similar alike the auditory ERB scale but the last 4 sub-bands differs from the ERB scale. However voice signals ranges upto 4,000 Hz and most of the speech energy lies below 1,500 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 are calculated by:

$$\langle S_i \rangle_k = \frac{\sum [w_\psi(x, k)_i]^2}{N_i} \quad (1)$$

where, $w_\psi(x, k)_i$ is the WP transform of signal x , i is the sub-band frequency index ($1 \leq i \leq M$), k represents the temporal frame and N_i is the number of coefficient in i th sub-band. The log of equal loudness weighted energy is calculated resulting a total of 24 coefficients. Discrete cosine transform (DCT) is applied on these 24 coefficients to de-correlate the filterbank energies and lower 13 coefficients are taken as

**Fig. 2** WERBC feature extraction technique

features. The block diagram of WERBC feature extraction is given in Fig. 2.

3 Robust feature extraction technique using SPADE

This section gives details of proposed SPADE based feature extraction technique. The detailed block diagram is given in Fig. 3. The development of this technique is strongly inspired by the outcomes of auditory comb filter hypothesis (Cheveigne' et al. 1997) and using of periodic and aperiodic feature in noisy speech recognition (Ishizuka and Miyazaki 2004; Ishizuka and Nakatani 2006). The hypothesis can be implemented in WERB-SPADE by using comb filters and ERB like WP decomposition of acoustic speech signal. This mechanism can be executed by decomposing speech signal into dominant periodicity and aperiodicity, which is the

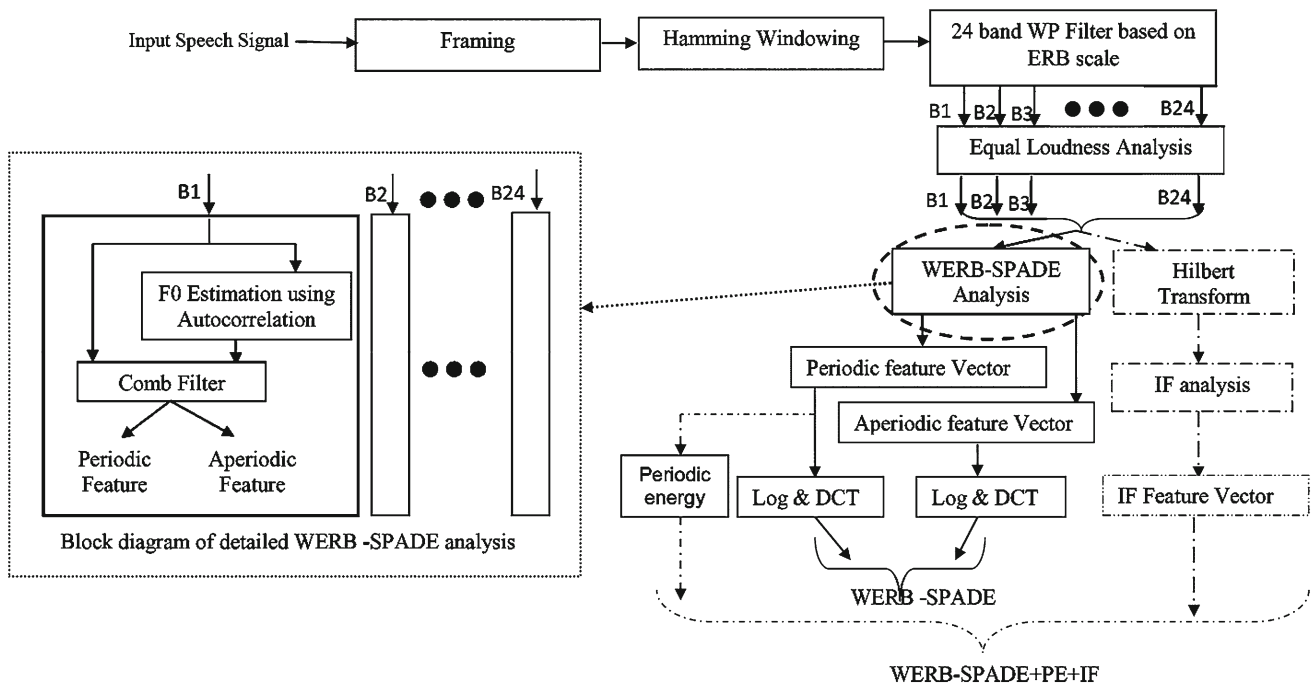


Fig. 3 Block diagram of WERB-SPADE analysis; shows the calculation of subband IF feature

residue power after the suppression of the dominant periodic power.

As similar to WERBC, hamming windowed frame size of 24 ms with 10 ms skip rate is used to derive the WERB-SPADE feature. In the next step, hamming windowed speech frame is decomposed into 24 wavelet sub-band as described in the Sect. 2. In the next step, output of each sub-band i.e. $w_\psi(x, k)_i$ (i th sub-band of k th temporal frame) is used to find dominant periodicity to design the comb filter. Dominant periodicity is calculated individually for each wavelet sub-band by applying a periodicity estimation technique such as the autocorrelation analysis for pitch (F0) estimation (Rabiner 1977). Comb filters are designed individually to take care of false periodicity detection in case of multi-pitch signals. It is expected SPADE mechanism can detect periodicity reliably. The basic function of a comb filter for each i th sub-band of k th frame can be expressed as:

$$h(n, k)_i = \sum_{l=-L}^L \alpha_l \times \delta(n - T_{i,k}) \tag{2}$$

where $\delta(n)$ is an unit impulse function, $(2L + 1)$ is the length of the filter, α_l is the filter coefficient satisfying $\sum_{l=-L}^L \alpha_l = 1$ and $T_{i,k}$ represents the dominant periodicity detected in each i th sub-band of k th frame. In the next step, the signal in the frame is passed through the comb filtered using the periodicity detected in the last step. Comb filter decomposes each sub-band signal into periodic and aperiodic

features. Comb filtering performed according to following equation:-

$$\langle w_\psi(x, k)_i \rangle_{comb} = h(n, k)_i \otimes (w_\psi(x, k)_i) \tag{3}$$

where $w_\psi(x, k)_i$ is the WP transform of signal x , i is the sub-band frequency index ($1 \leq i \leq M$), k represents the temporal frame and N_i is the number of coefficient in i th sub-band.

Next, the power suppressed by the comb filter is considered as periodic feature, and the residual signal power considered as aperiodic feature of speech signal. The calculation of periodic and aperiodic feature power vector is given below:

$$ap_i(k) = \frac{1}{N} \sum_{j=1}^N |\langle w_\psi(x, k)_i \rangle_{comb}|^2 \tag{4}$$

and,

$$p_i(k) = \frac{1}{N} \sum_{j=1}^N |w_\psi(x, k)_i|^2 - ap_i(k) \tag{5}$$

where, $p_i(k)$ and $ap_i(k)$ are periodic and aperiodic feature vector of i th sub-band of k th frame respectively, N is the number of confidents in i th sub-band. In the last step, the power vectors across the all sub-bands of the same frame is log transformed and cepstral coefficients is calculate using DCT.

$$c_r(k) = \sqrt{\frac{M}{2}} \sum_{i=1}^M \log(PV_i(k)) \cos\left(\frac{\pi r}{M}(i-0.5)\right) \quad (6)$$

where M is the number of sub-band; in our case 24, $PV_i(k)$ is the periodic power vector $p_i(k)$ or aperiodic power vector $ap_i(k)$ and $c_r(k)$ is the r th cepstral coefficient of k th temporal frame. Finally lower 13 coefficients from each power vector is taken and concatenated to single feature vector (26 features/frame) representing each temporal frame.

Additionally to make system more robust, periodic energy coefficient (PE) (Ishizuka and Nakatani 2006) is incorporated to basic WERB-SPADE feature vector. The simple log power of the sub-band coefficients can be easily affected by noise, but the log of periodic part of signal can have more resistant to noise. Periodic power energy is calculated by following formula.

$$E(k) = \sum_{i=1}^M p_i(k) \quad (7)$$

IF feature calculation is shown by dash lines in Fig. 3. The instantaneous frequency is calculated from the each wavelet sub-band output. By taking the derivative of phase of analytical band limited signal IF can be calculated efficiently. In this work we adopted the same technique to calculate IF as given by (Yin et al. 2011). The only difference is additional Hilbert transform is needed to calculate the analytical signal from WP sub-band outputs. IF is calculated for very WP sub-band, resulting a total of 24 IF features per frame.

4 Experimental framework

5 Speech corpus

The TIMIT corpus is used for all the experiments presented in this paper. TIMIT is one of the phonetically balanced standard corpus used to evaluate the performance of new techniques in ASR because 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 6,300 utterances for the eight major dialects of the United States. There are 630 different speakers, each one speaking ten sentences. For this work, dialect region DR1, DR2, DR3 and DR4 from training set is chosen for the extraction of phonemes from both male and female speakers. We mainly focus on the consonant (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/)) recognition because it is

one of most challenging task in ASR. The dialect region DR1, DR2, DR3 and DR4 from complete test set is used for testing. Furthermore, six noise such as car, jet, volvo, babble, speech, and lynx NoiseX-92 database is used in this work. Three kind of test set is used to evaluate the robustness of proposed technique. First one is test set A, contains clean test speech signal. Next speech signals from test set A mixed with car, jet and volvo noise at SNRs of -5 to 20 dB is named as test set B. Test set C contains speech signals mixed with babble, speech, and lynx noise at SNRs of -5 to 20 dB. The speech signals from dialect region DR5 to DR8 from complete test set is named as test set D. Next speech signals from test set D mixed with car, jet and volvo noise at SNRs of -5 dB to 20 dB is named as test set E.

In this experiment, two training set is used to train the Hidden Markov Model (HMM) (Lee and Hon 1989; Mes-saoud and Hamida 2010; Rabiner and Juang 1993; Young et al. 2009) based phoneme recognizer. First training set is called as clean-condition training set containing only clean speech signals. Another one is multi-condition training set containing clean speech signal and speech signal from test set B. Thus we have two types of training set and three types of testing set to study the performance of WERB-SPADE based techniques.

5.1 Experiment design

Six feature sets are extracted to study the performance of TIMIT phoneme recognition task. The methods adopted in this experiments are:-

- MFCC (Baseline): The feature set are derived using standard MFCC technique having 24 channel Mel scale filterbank. A frame size of 16 ms with 10 ms skip rate is used to analyze the speech signal. The 13 cepstral coefficients including energy coefficients are derived per speech frame.
- GFCC (Baseline): For GFCC the filter channel center frequencies are distributed according to ERB scale. 24 channel Gammatone filterbank is used to derive the features using same frame rate adopted for MFCC. After log compression and DCT operation first 13 features are taken per frame.
- WERBC: WERBC features are derived using db24 mother wavelet. Keeping same frame rate as baseline method 13 features are derived for each frame.

In addition, all features sets include the delta and acceleration coefficients and are 39-dimensional.

- WERB-SPADE: Wavelet ERB sub-band signals processed through SPADE analysis to decompose speech signal into periodic and aperiodic features resulting 26 fea-

Table 2 Average phoneme recognition accuracy with different features (without delta and acceleration coefficient)

Front end technique	Feature dimension/ frame	Clean condition training	Relative gain	Multi condition training	Relative gain
MFCC (Baseline)	13	72.37	NA	70.15	NA
GFCC (Baseline)	13	71.84	-0.73	71.27	1.60
WERBC	13	75.30	4.05	74.61	6.36
WERB-SPADE	26	78.38	8.30	77.58	10.59
WERB-SPADE +PE	27	80.55	11.30	80.38	14.58
WERB-SPADE +PE+IF	51	81.08	12.04	80.95	15.40

Table 3 Average phoneme recognition accuracy with different features (with delta and acceleration coefficient)

Front end technique	Feature dimension/ frame	Clean condition training	Relative gain	Multi condition training	Relative gain
MFCC (Baseline)	39	77.95	NA	75.60	NA
GFCC (Baseline)	39	76.24	-2.19	75.88	0.37
WERBC	39	82.67	6.06	81.96	8.41
WERB-SPADE	78	86.96	11.55	85.95	13.69
WERB-SPADE +PE	81	88.25	13.47	88.10	16.53
WERB-SPADE +PE+IF	84 (PCA)	88.90	14.05	88.84	17.51

tures per frame. Inclusion of delta and acceleration coefficients making it 78 features per frame.

- E. WERB-SPADE+PE: Periodic energy coefficient (PE) is included with WERB-SPADE to make 27 features per frame. Inclusion of delta and acceleration coefficients making it 81 features per frame.
- F. WERB-SPADE+PE+IF: 24 IF features are extracted across all 24 WP sub-band per temporal frame. Concatenating 24 features with WERB-SPADE+PE forms 51 features per frame. Inclusion of delta and acceleration coefficients making it 153 features per frame, which is too large in dimension in perspective of ASR. Thus principal component analysis (PCA) is applied to reduce the dimensionality of features. 84 PCA¹ transformed features per frame is taken finally.

One model is created for every phones and each HMM model has five emitting states. TIMIT phone-level annotation is used to create the phoneme model with 8 Gaussian mixture components with diagonal covariance, followed by Viterbi alignment to improve the state-time correspondence. The Baum-Welch algorithm is applied at the sentence level. Then, triphone context dependent (CD) HMMs (Lee and Hon 1989) are created using the pronunciation lexicon. Once compiled, the Viterbi decoder with whole recognition network can be used to classify the phoneme into their respective classes for an unknown input utterance. Phoneme

recognition accuracy (PRA) is calculated by the following equations:

$$PRA(\%) = 100(\%) - PER(\%) \quad (8)$$

where Phoneme error rate (PER) is given by:-

$$PER(\%) = \frac{(Substitutions + Deletions + Insertions)}{Total\ Phoneme} \times 100 \quad (9)$$

$$= Subs(\%) + Del(\%) + Ins(\%)$$

6 Results and discussions

We have started the experiment in sequential steps started with clean test data followed by noisy condition.

6.1 Performance evaluation under clean condition

Initially, CD phone recognition task is carried out with features which does not consider contextual information (delta and acceleration coefficients) appended in the feature vector. Table 2 shows the average phone recognition accuracy (Test set A) of all six methods under clean-condition training and multi-condition training. Relative percentage gain compared to the commonly used MFCC features are also mentioned in Table 2. Table 3 shows the performance analysis with delta and acceleration analysis. As expected, systems trained with dynamic features shows better performance because it captures the dynamic informations presents in speech signal. From Table 2 and 3, it is exciting to see while comparing

¹ PCA dimension applied in the range of 40–90, Optimal PCA transformed features are used.

two baseline methods, auditory based GFCC features performed worst compared to MFCC in clean condition. But in multi-training condition performance of MFCC is dropped significantly compared to GFCC features.

MFCC has poor recognition efficiency in case of unvoiced phonemes and seems more vulnerable in presence of noise. Due to this there is some inadequacies in clean condition training using MFCC. 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 scale while working in noisy condition. Further WERBC feature outperformed baseline features because it has benefit of time-frequency wavelet analysis along with sub-band decomposition according to the ERB scale. As described in Sect. 2 we have focused 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 the lower frequency region where more perception (frequency discrimination) is available.

Now we change the focus on wavelet SPADE based techniques. WERB-SPADE outperformed WERBC features by a significant margin in every case. SPADE estimates periodicity reliably in each sub-band to successfully reveal the two important property of the speech signals. Inclusion of aperiodic features helps to improve to recognize unvoiced phonemes such as stops, fricative and nasals. Thus without losing the certain information about aperiodicity of signal, WERB-SPADE provides more robustness compared to WERBC. WERBC decomposed the speech signals in the number of sub-bands according to the ERB scale by utilizing the rich covering of time-frequency property that can enrich the features to represent the short term behavior of the speech in well manner compared to STFT based techniques. WP provides an effective way to produce sub-band dependent partitions of the observation space and representing each sub-band into periodic and aperiodic speech features through comb filtering helps to catch essential discriminative information embedded in speech signal in an effective way. Further inclusion of PE coefficient in WERB-SPADE increases the performance by significant margin. From both table, it is interesting to see switching the scenario from clean train to multi-train condition having negligible effect on performance of phone recognizer compared to above mentioned techniques. Periodic signal power in speech considered to be more robust to noise than the simple power until and unless the interfering noise signal does not having stronger periodicity than the target speech. Thus despite of inclusion of noisy speech from test set B not having significant impact on Gaussian mixture during multi-condition training with

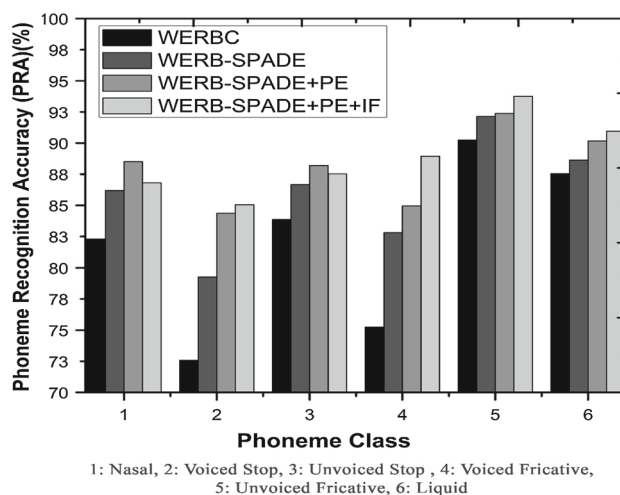


Fig. 4 Detailed phone recognition accuracy with wavelet based features

WERB-SPADE +PE features. Next, we changed the focus on IF based WERB-SPADE, which has not shown significant improvement over WERB-SPADE +PE with clean test set A. Although it has shown best recognition efficiency among all techniques despite of 44.3 % dimension reduction. Thus considering phase information derived from the output of each wavelet sub-band along with WERB-SPADE +PE has proved their discriminative property in both training conditions. From now on for better analysis of performance of wavelet based techniques we consider the multi-training condition system trained with dynamic features. Detailed phoneme recognition is shown in Fig. 4 and extensive PER analysis of wavelet SPADE based techniques is demonstrated in Table 4. From Fig. 4 it can be noted that the performance of WERBC is significantly goes low with voiced phonemes especially stops. Voiced consonants are periodic in nature and WP decomposition might be less sufficient to extract the periodic structure embedded in speech signal. Another reason of this low recognition rate might be explained by the fact that the burst spectrum arises in voiced stop and its voiceless counterpart is very similar (Niyogi and Ramesh 2003) causing misclassification towards unvoiced stop. The recognition performance of voiced phoneme is enhanced with wavelet SPADE based recognizer because it finely represents the speech by its periodic and aperiodic property. Thus periodic information embedded in voiced phonemes is captured efficiently in WERB-SPADE based techniques

6.2 Performance evaluation in dialect mismatch condition

To study the effectiveness of spade based wavelet feature in dialect mismatch condition test set D and test set is used. Table 5 has shown the recognition performance of all features in multi-training condition. Table 5 also shows the relative changes in performance of dialectal mismatch con-

Table 4 Detailed PER analysis with different types of error

	Deletion			Intra substitution			Inter substitution			Insertion		
	Feature D	Feature E	Feature F	Feature D	Feature E	Feature F	Feature D	Feature E	Feature F	Feature D	Feature E	Feature F
Nasals	3.05	2.25	2.74	5.78	4.86	5.02	3.65	3.35	3.70	1.32	1.02	1.72
Voiced stop	4.14	2.87	3.00	7.35	5.47	5.28	5.20	4.12	3.41	4.04	3.17	3.26
Unvoiced stop	3.05	2.74	2.85	5.45	4.98	4.72	1.95	2.40	2.20	2.89	1.68	2.71
Voiced fricative	2.79	3.14	2.50	6.90	5.91	4.80	5.75	3.25	2.94	1.76	2.75	0.81
Unvoiced fricative	2.05	1.55	0.88	3.48	2.54	2.25	1.85	1.78	2.08	0.49	1.75	1.04
Liquids	2.24	2.95	2.15	4.21	3.47	3.82	3.43	2.87	2.35	1.48	0.53	0.73
Avg	2.89	2.58	2.35	5.53	4.54	4.32	3.64	2.96	2.78	2.00	1.82	1.71

Table 5 Performance evaluation of dialect mismatch condition Relative change (%) is shown in parentheses compared to performance of non-dialect mismatch condition

Feature extraction technique	Test set D		Test set E				
	Clean	20 dB	15 dB	10 dB	5 dB	0 dB	(−)5dB
MFCC (baseline)	72.74(−3.79 %)	68.39(−5.78 %)	65.14(−3.51 %)	57.87(−2.86 %)	48.10(−5.06 %)	36.21(−12.84 %)	25.30(−7.13 %)
GFCC (baseline)	72.59(−4.34 %)	69.03(−7.81 %)	66.02(−9.28 %)	60.86(−3.83 %)	54.92(−3.07 %)	43.49 (−3.51 %)	29.21(−3.32 %)
WERBC	80.62(−1.64 %)	77.10(−0.83 %)	72.75(−1.84 %)	64.20 (−3.6 %)	55.82(−3.47 %)	46.92 (−2.27 %)	31.87(−4.79 %)
WERB-SPADE	85.31(−0.75 %)	79.30(−0.88 %)	76.14(−1.12 %)	71.24(−1.39 %)	59.10(−2.29 %)	48.55 (−2.02 %)	34.20(−2.85 %)
WERB-SPADE +PE	87.26(−0.96 %)	79.68(−0.58 %)	76.10(−0.72 %)	74.05(−1.21 %)	63.82(−2.18 %)	52.31 (−1.77 %)	36.97(−2.07 %)
WERB-SPADE +PE+IF	88.35(−0.56 %)	81.18(−0.45 %)	79.23(−1.55 %)	77.13(−1.71 %)	63.17 (0.05 %)	55.41 (−1.43 %)	38.42(−2.67 %)

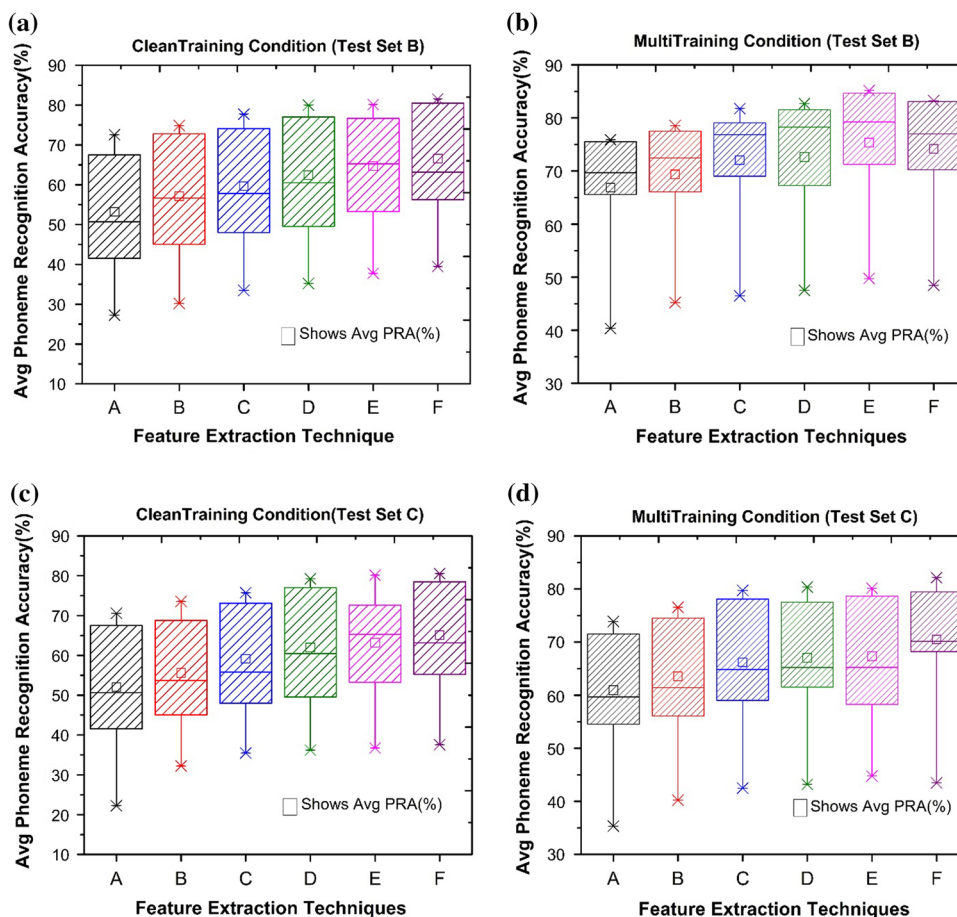
dition (Test set D and E) and dialect non-mismatch condition (Test set A and B). The performance of MFCC features drops down relatively by significant margin due to the fact that Mel scale might be less superior to track the dialectal changes which slows down phoneme recognition. Compared to MFCC, GFCC features have shown significant robustness at low SNRs in dialectal mismatch condition which shows the adaptability of ERB to the dialect mismatch condition. WERBC features have performed better than GFCC by use of time frequency analysis property of WP. However, by emphasizing the periodicity and aperiodicity analysis on individual wavelet sub-band, SPADE based WERBC features have shown better recognition efficiency.

7 Performance evaluation under noisy condition

The experimental results using different front end features with noisy speech (Test set B and Test set C) and under clean and multi-condition training are shown in Fig. 5. This clearly shows the improved performance of the WP derived features for English phonemes over MFCC and GFCC features. WP

derived features are less sensitive to interfering signal and can focus to the target speech in composite auditory scene to extract the coefficients at a certain frequency of interest. Further, the results clearly has shown the effectiveness of WERB-SPADE based features in the presence of noise. These methods could considerably enhance noise robustness by using the periodic information of the speech signal, because periodicity is essentially less affected by interfering signal. The decomposition of WP sub-band into two feature set helps to reduce the impact of the signal power distortion in noisy environments. In addition, aperiodic features clearly reflects sound onset and frequency transitions in a frame (Ishizuka and Miyazaki 2004). Further, the inclusion of pitch in the WERB-SPADE further increased the robustness of the ASR system. On average the best performance is achieved while considering the IF features. Fig 5(b) shows WERB-SPADE +PE+IF is outperformed by WERB-SPADE +PE with a relative margin of 1.6 % in multi-training condition with test set B while scenario is reversed in case of test set C by getting ample relative improvement of 4.67 %. By considering the IF features perhaps reduce the environmental mismatch between training and testing data. By capturing

Fig. 5 Average recognition accuracy of all phoneme classes under SNRs of -5 to 20 dB. Lower whisker shows the average recognition efficiency under -5 dB while upper whisker represents 20 dB. (Color figure online)



the phase information as a function of frequency concatenated IF-based features shows robust discriminative property in noisy environment even in environmental mismatch condition.

8 Conclusion

This paper proposed a robust front end WP based feature extraction method which capture the periodic information embedded in speech without losing important aperiodic information. The method uses WP sub-band decomposing using auditory ERB scale and each sub-band is processed through comb filter to derive periodic and aperiodic feature set. TIMIT phone recognition task is carried out to study the performance of proposed feature. WERB-SPADE is found to be more effective to recognize the phoneme compared to baseline features. Further PE is included to make the feature more robust against noise. Inclusion of IF features make the WERB-SPADE more effective in environmental mismatch condition as seen in presented results. Although proposed feature exhibits effectiveness in the presence of noise but these features have larger number of parameter compared to baseline methods and WERBC. So computational cost could

be an important factor in real time speech processing. This is important to find out a trade-off between feature dimension and performance to make it more effective in real time conditions.

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.

References

- Adami, A., Burget, L., Duponi, S., Garudadri, H., Grezl, F., Hermansky, H., Jain, P., Kajarekar, S., Morgan, N. & Sivasdas, S. (2002). QUALCOMM-ICSI-OGI features for ASR. Proceedings ICSLP. pp 21–24.
- Ali, A. M., Spiegel, J. V., & Mueller, P. (2002). Robust auditorybased speech processing using the average localized synchrony detection. *IEEE Transactions on Speech and Audio Processing*, 10, 279–292.
- Berouti, M., Schwartz, R. & Makhoul, J. (1979). Enhancement of speech corrupted by acoustical noise. Proceedings ICASSP. pp 208–211.
- Biswas, A., Sahu, P. K., & Chandra, M. (2014). Admissible wavelet packet features based on human inner ear frequency response for Hindi consonant recognition. *Computer and Electrical Engineering*, 40, 1111–1122.
- Cheveigne, D., McAdams, A., & Marin, S. (1997). Concurrent vowel identification. II. Effects of phase, harmonicity, and task. *The Journal of the Acoustical Society of America*, 101, 2848–2856.

- Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics Speech and Signal Processing ASSP*, 28, 357–366.
- El-Fattah, M.A.A., Dessouky, M.I., Abbas, A.M., Diab, S.M., El-Sayed, M., El-Rabaie, W.A.-N., Alshebeili, S.A., El-samie, F.E.A. (2013). Speech enhancement with an adaptive Wiener filter. *Int J Speech Technol*, 1–12.
- Farooq, O., & Datta, S. (2001). Mel filter-like admissible wavelet packet structure for speech recognition. *The IEEE Signal Processing Letters*, 8, 196–198.
- Farooq, O., & Datta, S. (2004). Wavelet based robust sub-band features for phoneme recognition. *IEE Proceedings-Vision, Image and Signal Processing*, 151, 187–193.
- Farooq, O., Datta, S., & Shrotriya, M. C. (2010). Wavelet sub-band based temporal features for robust Hindi phoneme recognition. *International Journal of Wavelets, Multiresolution and Information Processing*, 8, 847–859.
- Gao, Y., Huang, T., Chen, S., Hatan, J. (1992). Auditory model based speech processing. *Proceedings ICSLP*. pp 73–76.
- Garofolo, J.S., Lamel, L.F., Fisher, W.M., Fiscus, J.G. & Pallett, D.S. (1993). DARPA TIMIT acoustic-phonetic continuous speech corpus CD-ROM. NIST speech disc 1–1.1. NASA STI/Recon Technical Report N, 93, 27403.
- Greenberg, S., Ainsworth, W. A., Popper, A. N., & Fay, R. R. (2004). *Speech processing in the auditory system: an overview*. New York: Springer-Verlag.
- Hermansky, H. (1990). Perceptual linear predictive (PLP) analysis. *Journal of the Acoustical Society of America*, 87, 1738–1752.
- Hohmann, V. & Kollmeier, B. (2006). A nonlinear auditory filterbank controlled by sub-band instantaneous frequency estimates. *Int. Symp. Hear. 2006* (Springer). Cloppenburg, Germany, pp 11–18.
- Ishizuka, K. & Miyazaki, N. (2004). Speech feature extraction method representing periodicity and aperiodicity in sub bands for robust speech recognition. *Proceedings ICASSP*. pp 141–144.
- Ishizuka, K., & Nakatani, T. (2006). A feature extraction method using sub-band based periodicity and aperiodicity decomposition with noise robust frontend processing for automatic speech recognition. *Speech Communication*, 48, 1447–1457.
- Kajita, S. & Itakura, F. (1995). Robust feature extraction using SBCOR analysis. *Proceedings ICASSP*. pp 421–424.
- Kim, D. S., Lee, S. Y., & Kil, R. M. (1999). Auditory processing of speech signals for robust speech recognition in real-world noisy environments. *IEEE Transactions on Speech and Audio Processing*, 7, 55–69.
- Lee, K. F., & Hon, H. W. (1989). Speaker-independent phone recognition using hidden Markov models. *IEEE Transactions on Acoustics Speech and Signal Processing*, 37, 1641–1648.
- Li Q., Soong, F.K. & Siohan, O. (2001). An auditory system-based feature for robust speech recognition. *Proceedings Eurospeech*. pp 619–621.
- Long, C. & Datta, S. (1996). Wavelet based feature extraction for phoneme recognition. *4th Int. Conf. Spok. Lang. Process. Philadelphia (USA)*, pp 264–267.
- Mallat, S. (1989). A theory for multiresolution signal decomposition: the wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 674–693.
- Messaoud, Z. B., & Hamida, A. B. (2010). Combining formant frequency based on variable order LPC coding with acoustic features for TIMIT phone recognition. *International Journal of Speech Technology, Springer*, 14, 393–403.
- Niyogi, P., & Ramesh, P. (2003). The voicing feature for stop consonants: recognition experiments with continuously spoken alphabets. *Speech Communication*, 41, 349–367.
- Patterson, R. D., Nimmo-Smith, I., Holdsworth, J., & Rice, P. (1988). *An efficient auditory filterbank based on the Gammatone function*. Cambridge: Appl. Psychol. Unit, Cambridge University.
- Rabiner, L. R. (1977). On the use of autocorrelation analysis for pitch detection. *IEEE Transactions on Acoustics Speech and Signal Processing*, 25, 24–33.
- Rabiner, L. R., & Juang, B. H. (1993). *Fundamentals of speech recognition*. Englewood Cliffs, USA: PTR Prentice-Hall.
- Shao, Y., Srinivasan, S., Jin, Z., & Wang, D. (2010). A computational auditory scene analysis system for speech segregation and robust speech recognition. *Computer Speech & Language, Elsevier*, 24, 77–93.
- Stark, A.P. & Paliwal, K.K. (2008). Speech analysis using instantaneous frequency deviation. *the Interspeech*. Brisbane, Australia, pp 2602–2605.
- Yin, H., Hohmann, V., & Nadeu, C. (2011). Acoustic features for speech recognition based on Gammatone filterbank and instantaneous frequency. *Speech Communication*, 53, 707–715.
- Young, S.J., Evermann, G., Gales, M.J.F., Hain, T., Kershaw, D., Liu, X., Moore, G., Odell, J., Ollason, D., Povey, D., Valtchev, V. & Woodland, P.C. (2009). *The HTK Book* (for HTK Version 3.4). Construction 384.