Int J Speech Technol (2016) 19:881–893
DOI 10.1007/s10772-016-9378-9

Simulation and overall comparative evaluation of performance between different techniques for high band feature extraction based on artificial bandwidth extension of speech over proposed global system for mobile full rate narrow band coder

Ninad S. Bhatt[1]

Received: 9 February 2016 / Accepted: 4 October 2016 / Published online: 8 October 2016

Abstract This paper addresses a novel approach to investigate, study and simulate computation of high band (HB) feature extraction based on linear predictive coding (LPC) and mel frequency cepstral coefficient (MFCC) techniques. Further, HB features are embedded into encoded bitstream of proposed global system for mobile (GSM) full rate (FR) 06.10 coder using joint source coding and data hiding before being transmitted to receiving terminal. At receiver, HB features are extracted to reproduce HB portion of speech and for the same different extension of excitation techniques are applied and their results evaluated in terms of quality (intelligibility and naturalness) and bandwidth. MATLAB based e-test bench is created for implementing the proposed artificial bandwidth extension (ABE) coder following series of simulations, that are carried out to discover and gain insight about the performance of it using subjective [mean opinion score (MOS)] and objective [perceptual evaluation of speech quality (PESQ)] analysis. The results obtained for both the analyses advocate that proposed ABE coder outperforms proposed GSM FR NB (legacy GSM FR) coder. While the fact remains that, compared to LPC based parameterizations over ABE coder, MFCC parameterization results in higher speech intelligibility which is evident from obtained slightly better PESQ and MOS scores.

Keywords GSM full rate coder · Artificial bandwidth extension · Linear predictive coding · Mel frequency cepstral coefficients · Subjective analysis · Objective analysis

✉ Ninad S. Bhatt
bhattninad@gmail.com

[1] Department of Electronics and Communication Engineering, C. K. Pithawalla College of Engineering & Technology, Surat, Gujarat, India

1 Introduction

Present underlying wireless communication system inherently suffers from limited acoustic narrow bandwidth of 300–3400 Hz for speech communication. In comparison with conventional narrow band (NB) systems, wide band (WB) speech (having cut-off frequency $f_c = 7$ kHz) could considered to be desirable because it guarantees significant increase in subjective speech quality, intelligibility and also reduces listening efforts (Bhatt et al. 2012; Jax and Vary 2006). Along with limited acoustic bandwidth, other factors which may affect the quality of recovered speech are acoustic background noise, quantization noise due to source coding and residual error after channel decoding. Fricatives like /s/, /z/ and partly /f/, /S/, /Z/ are difficult to recover using only NB speech as considerable energy of these fricatives are located in high frequency band thus limiting the performance and speech quality of recovered speech signal (Bhatt et al. 2012).

Since inception, aggressive researches in the emerging era of WB speech coding have been reported by both researchers and academic communities. As an outcome, many WB coders have been evolved and standardized by European Telecommunication Standards Institution (ETSI) and International Telecommunication Union-Terrestrial (ITU-T) organizations, and, also found quite superior in terms of substantial quality improvement in recovered speech in context with former. The major obstacle in deployment of WB coders and WB wireless systems is backward compatibility. As WB systems need WB end terminals and WB supporting network infrastructure, which may ultimately lead to up-gradation of entire NB system and that may incur huge investment expenses to network service providers and eventually end up to pockets of customers. Thus a long transition time elapsed in full

functioning and deployment of WB systems and WB coders, demanded scope of development of competent potential candidate like ABE that meets the requirement of fulfillment of WB comparable recovered quality of speech over existing NB transmission network setups (Jax and Vary 2003, 2006).

This research aims at investigating a novel approach on state-of-the-art ABE coder based on LPC and MFCC techniques for computing HB features; representing HB portion of WB speech (frame wise i.e. 36 bits/20 ms frame). The same can then be encoded using non uniform quantization and such formed coded bitstream could be embedded into bitstream of proposed GSM FR NB coder (224 bits/20 ms frame) using joint source coding and data masking/hiding techniques. This effectively produces 260 bits/20 ms frame as per GSM 06.10 FR coder standardized by ETSI (2005-06). HB features, recovered at receiver by watermark extraction section, can be used to determine HB speech. Salient feature of this work is to deploy and examine numerous extension of excitation techniques for HB speech recovery. NB version of recovered speech could be an outlet from proposed GSM FR NB (legacy GSM FR) decoder. Both NB and HB version of speech are then summed up using over lap add method to generate WB recovered speech signal at receiving end.

The remainder of paper is organized as follows: Overview of proposed GSM FR 06.10 NB coder is outlined in Sect. 2. Section 3 focuses on adoption of ABE coder based on LPC & MFCC feature extraction techniques; followed by demonstration and examination of overall performance comparison analysis and results of proposed ABE coder in Sect. 4. Finally, concluding remarks are narrated in Sect. 5.

2 Proposed modifications in GSM FR 06.10 NB coder

ETSI GSM 06.10 Full Rate Coder is classified into three major folds: Linear predictive coding (LPC) section, long term predictive (LTP) section and regular pulse excitation (RPE) section (ETSI 2005-06). The proposed modifications are suggested in RPE section in which the selection procedure of grid position and samples is modified in such a fashion that no samples repeat in multiple grids which is the case of ETSI standard GSM Full Rate coder (first and forth grid where except sample number 0 and sample number 39 don't repeat where as all other samples in both grids repeat). A newly proposed grid selection strategy is highlighted in Fig. 1 where if the weighting filtered prediction-error sequence is down-sampled by a ratio of 4 instead of 3, it results into four interleaved sequences with

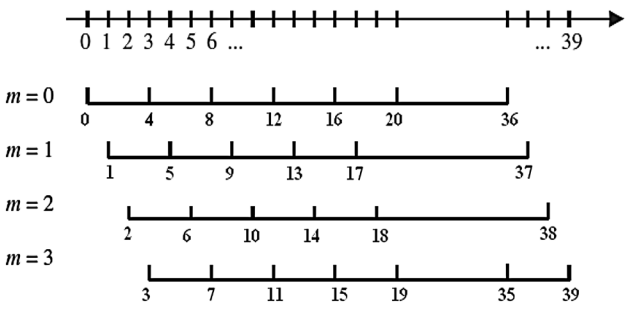

Fig. 1 Sampling grids used in position selection for proposed GSM FR 11.2 kbps coder (Bhatt and Kosta 2011)

regularly spaced pulses (Bhatt and Kosta 2011). These are mathematically expressed as follows

$$x_m(k) = x(m + 4k) \quad (1)$$

m = 0, 1, 2, 3; k = 0, 1, 2…9;
where, m = no. of grids per sub segment and k = no. of samples per grid.

The major merit in this strategy of sampling grid position selection is; no repetition of any sample in multiple grids, at the same time, total number of samples per grid reduces from 13 to 10 so effectively there is a reduction in overall bit-rate of 1.8 kbps (3 samples per grid * 3 bits per sample * 4 sub frames = 36 bits/20 ms frame) compared to actual bit rates of 13 kbps for ETSI standard GSM 06.10 FR coder (Bhatt and Kosta 2011). The proposed modification in GSM FR offers a new bit allocation as illustrated in Table 1.

Primary aspect of the proposed modification in GSM FR coder is to spare and utilize 1.8 kbps bitstream (36 bits per each frame) for HB feature transmission in class I_b as per Table 3 of Bhatt et al. (2011) as per channel coding standards described in ETSI 05.03 (1999).

3 Implementation of proposed ABE coder

Distinguish feature of this implementation of proposed ABE coder is its adoption of HB side information extraction (from HB speech) based on LPC and MFCC techniques; embedding and transmitting HB information over proposed GSM FR NB coder. At receiver, along with accomplishment of NB speech recovery, HB speech regeneration process is conducted by integrating several extension of excitation techniques like spectral folding (SF), spectral translation (ST), non linear distortion (NLD), noise modulation (NM), full wave rectification (FWR) and sinusoidal transform coding (STC).

Though, there exist various approaches that claim to represent the feature vector of HB speech, candidature of LPC and MFCC are considered (Nour-Eldin and Kabal 2008) with following reasons:

Table 1 Bit allocation for proposed GSM full rate speech coder (Bhatt and Kosta 2011)

Parameter	No. per frame	Resolution	Total bits/frame
LPC	8	6,6,5,5,4,4,3,3	36
Pitch Period	4	7	28
Long Term Gain	4	3	12
Grid Position	4	2	8
Peak Magnitude	4	6	24
Sample Amplitude	4*10	3	120
Total	224		

- Computation of both of the above approaches leads to source filter separation
- Both of LPC and MFCC parameters have analytically tractable models
- Importance of both the approaches in speech recognition applications

3.1 Implementation of LPC based proposed ABE coder

This subsection demonstrates detailed framework of LPC based proposed ABE coder. Entire process flow of feature extraction, embedding cum data masking/hiding over NB bitstream and recovery of WB speech via exploiting different excitation generation has been studied, investigated and simulated.

Band splitting and recombining is considered to be the process which is carried out in initial and final phase of proposed ABE coder. As depicted in Fig. 2, having applied WB speech corpus (having bandwidth up to 7 kHz and sampling frequency of 16 kHz) as an input to proposed ABE coder; further, splitting of WB signal into NB and HB signal can be yielded using the LPF and HPF respectively (both having cut-off freq. equal to 3.4 kHz) followed by down sampling both of them with the factor of two. As standard ETSI GSM FR coder contains time frame each having 160 samples per 20 ms resulting into effectively 260 bits per each frame (ETSI 2005-06); frame wise NB speech has been processed by proposed GSM FR coder resulting into encoded bitstream of 224 bits per frame (11.2 kbps) (Bhatt and Kosta 2011). Watermark embedding and masking serves to hide side information bits (1.8 kbps) into bitstream of proposed GSM FR coder before sending over NB wireless link at a standard bitrates of 13 kbps.

At receiver, NB synthetic speech can be reproduced using proposed GSM FR NB decoder and up-sampled by factor of two which is then added using Over Lap Add method with synthetically generated HB version of speech to finally regenerate synthetic WB speech signal.

HB feature extraction block computes HB feature representing side information bits (36 bits per frame) that are being determined based on 8^{th} order LPC technique from a given HB version of speech. In order to derive LP coefficients ($a_0 \ldots a_7$), auto correlation followed by Levinson Durbin algorithm is conducted (Jax et al. 2006; Geiser and Vary 2007). Also, gain parameter per frame is computed from the given HB signal. Thus, eventually LP coefficients act as representative candidate parameter of HB features side information on frame wise basis; over which non-uniform quantization is being adopted depending upon the subjective importance of individual parameter.

Table 2 illustrates bit allocation of LP coefficients as well as gain parameter.

Main ideology of the proposed ABE coder is to yield HB excitations at receiving end only from NB excitations itself utilizing different extension of excitation techniques. So, in stark contrast to the other conventional approaches, the proposed method thus eliminates the basic need of transmitting HB excitations (which may include pitch period information etc.) to receiver; in turn reduces

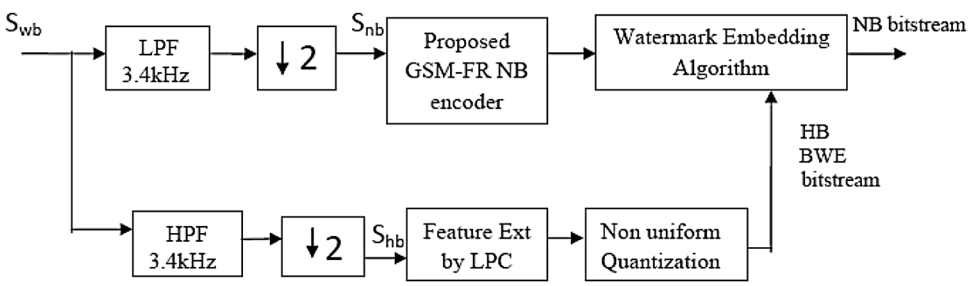

Fig. 2 Transmitter of proposed BWE coder based on LP coefficients

Table 2 Non uniform quantization and bit allocation of LPC coefficients representing HB features

Parameter	No. per frame	Resolution	Total bits/frame
LPC coefficients	08	6,5,5,4,4,3,3,3	33
Gain	01	3	3
Total	36 bits		

overhead bits required for representing spectral envelop at receiver. It remains fact that lesser the overhead bits lesser the distortion offered because of embedding of HB feature information on the NB coded speech.

Watermark embedding and data hiding section highlights that in the recent past decade lot many data hiding techniques have been evolved for hiding information over career of speech signals. Such techniques are least significant bit (LSB) insertion, spread spectrum (SS), quantization index modulation (QIM), auditory masking (AM), phase coding (PC), Echo Coding (EC) etc. (Shahbazi 2010; Bhatt et al. 2011). In principle, there exist three basic classifications of steganographically data hiding over NB channels; signal domain data hiding, bitstream data hiding and joint source coding cum data hiding (Vary and Geiser 2007).

The watermark embedding process addressed in this research can be distinguished in comparison with other techniques depicted in Vary and Geiser (2007) in a way that the proposed coder aims at provisioning room (space) for data hiding simultaneously while performing source coding which in this case referred to proposed GSM FR encoding. As discussed in previous subsection and also demonstrated in Table 3 of Bhatt et al. (2011), for proposed GSM FR coder, initially class I_b contains 96 bits (bit location d50–d145) and the quantized bits of HB information, as shown in Table 2, are embedded at the bit location of d146–d181 (36 bits of class I_b) which finally generates 260 bits/frame as per standard ETSI GSM FR coder (2005-06). It should be brought to notice that proposed GSM FR coder produces 224 bits/frame and other 36 bits of quantized HB feature bits (side information) are embedded by watermark embedding algorithm using joint embedding and data hiding approach. Primary goal of watermark embedding and hiding in this work is to embed and transmit HB features over GSM NB bitstream; hence no other steganographic techniques are touched upon or focused for securing the embedded data itself. Moreover, as HB features are added in class I_b, channel coding while exploring convolution coding can be sufficient for error protection of stego bitstream.

WB speech regeneration process at receiving end is demonstrated in Fig. 3, in which receiving end accepts NB bitstream on the frame wise basis (260 bits/frame). Watermark extraction algorithm is responsible to separate out 36 bits [from class I_b of Table 3 of Bhatt et al. (2011)] for HB speech recovery and pass on the bitstream of 224 bits/frame to proposed GSM FR NB decoder (legacy decoder) for decoding and reproduction of NB version of speech. The produced NB speech can then be up-sampled by the factor of two and subsequently passed through LPF having cut-off frequency of 3.4 kHz to ultimately regenerate final NB speech. Frame wise separated 36 bit coded bitstream of HB features by watermark extraction are then dequantized to produce LP coefficients. It is worth noting that in order to generate NB excitation signal, NB decoded speech should be processed through LP analysis filter first (using HB features) before performing extension of excitation in only those techniques which don't have provision of inbuilt analysis filtering. WB excitation signal is produced from the available NB signal (generated from proposed GSM FR decoder) using different extension of excitation techniques discussed below. Along with generated WB excitation, the reproduced features are interpolated by the factor of two (extension of envelop) and then supplied to their respective blocks like gain adjustment and synthesis filter to effectively reconstruct WB version of speech. Further, WB speech is processed through HPF having cut-off frequency of 3.4 kHz to yield HB speech recovery. Finally, both NB and HB recovered signals are

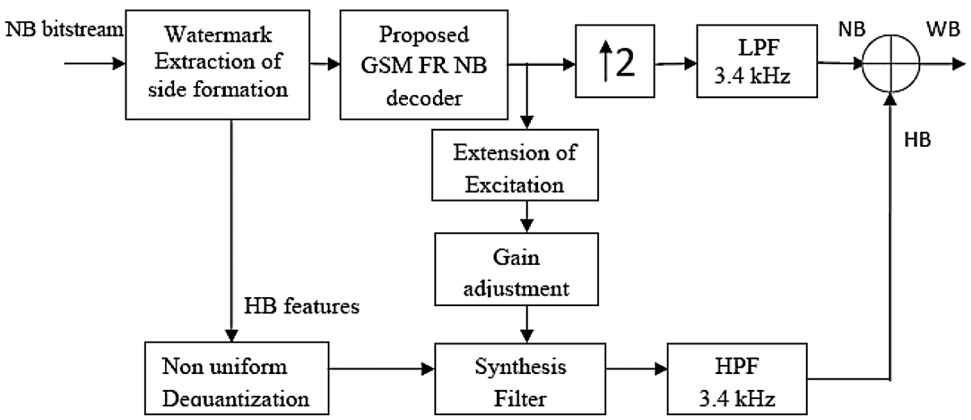

Fig. 3 Receiver section of LPC based proposed BWE coder

summed up using over lap add method with length equalization to reconstruct WB speech at receiver.

Extension of excitation process is an important aspect of any wideband enhancement scheme is the generation of the HB/WB speech excitation. The residual extension process mainly claims to double the sampling rate, from 8 to 16 kHz, as well as to maintain the whole spectrum flat. The harmonics contained in the NB residual should also be carried forwarded in the WB residual. Forthcoming are few of the popular techniques which are referred in this research for bringing forth HB/WB excitation signal from the given input NB excitations (NB residual signal).

- Spectral folding (SF)
- Spectral translation (ST)
- Non linear distortion (NLD)
- Noise modulation (NM)
- Full wave rectification (FWR)
- Sinusoidal transform coding (STC)

Following is the description cum overview of the techniques for WB excitation generation which are implemented in this work and the performance of proposed ABE coder for these techniques are being studied, simulated and analyzed in this research.

Spectral folding is basically a time-domain approach for extension of excitation and is one of the most conventional and popular methods because of its simplicity and wide usage in high frequency excitation regeneration in ABE. Inserting zero between each sample of NB residual signal, folds the baseband spectrum to higher frequencies. Thus, the up-sampling extends the signal bandwidth from 4 to 8 kHz. Spectrum folding results in the mirrored image of the NB spectrum at the spectrum of HB but the major disadvantage of this method is that the harmonic structure leaves a spectral gap at 4 kHz (Fuemmeler et al. 2001; Cabaral and Oliveira 2005).

Spectral translation creates a shifted spectral version of the NB residual at the high frequency band. This approach prompts input NB residual to mix with $(-1)^n$, (where n is the index of each sample) which is rather similar to modulation of any signal with frequency equal to Nyquist frequency. This modulated signal is up-sampled and high pass filtered and added to the up-sampled and low pass filtered NB residual to yield WB residual at the output. Major shortfall of this approach is that it fails to preserve the original NB residual information (Fuemmeler et al. 2001)

In the *non linear distortion* approach, initially NB residual is up-sampled by factor two and then delivered to a nonlinear function as expressed in Eq. 2. The desired bandwidth and harmonic structure can be obtained over the whole spectrum from the resulted distorted signal. After the whitening filter, the resulting signal spectrum is then flattened so that the excitation does not affect the overall spectral shape. Finally, the output meets the requirement of WB residual. A simple nonlinear function can be illustrated by:

$$y(t) = [(1+\alpha)\,|x(t)| + (1-\alpha)x(t)]/2 \tag{2}$$

where, $x(t)$ is the input signal, $y(t)$ is the distorted output signal, and α is a parameter between 0 and 1. When $\alpha = 1$, it becomes the absolute value function. Furthermore, after varying the values of α within a specific stipulated range on the trial and error basis in this work, it is witnessed that value of $\alpha = 0.7$ offers overall promising and better results (Cabaral and Oliveira 2005).

Noise modulation is an exemplary approach for residual excitation extension, because according to the human auditory system perceiving capabilities of human ear gradually reduces with the increase in the frequency above 4 kHz. Even the harmonic structure of speech signal itself is lost at higher frequencies and becomes more noise-like. This put forwards that in HB speech model, the 3–4 kHz band of NB speech is made use of bringing out its time domain envelope and the HB excitation is produced by modulating HB noise using this envelope. Figure 4

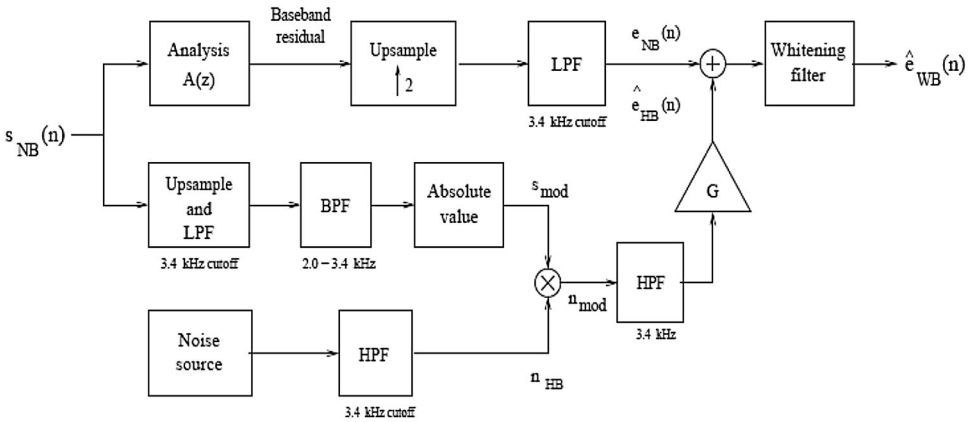

Fig. 4 Noise modulation technique

illustrates the process of noise modulation for HB excitation generation. For the same, white gaussian noise is generated with a positive random sequence having mean equal to 2 and variance of 0.5. This modulated signal is subsequently high pass filtered with cut-off frequency of 3.4 kHz and also the energy is scaled of the resulting HB excitation signal by NB excitation. Further, the up-sampled and low pass filtered NB residual is added in the high frequency modulated noise i.e. HB excitation to regenerate WB excitation. Finally, WB excitation is processed through the whitening filter to have flat spectrum of resulting excitation signal (Cabaral and Oliveira 2005).

Though, Noise Modulation is computationally efficient and the same is also motivated by human perception, but it has inherent drawback of being dependent on the assumption that the time domain envelope of the 3–4 kHz speech band is identical to that of the 4–8 kHz band.

Full wave rectification approach is also a step forward towards demonstrating how wideband excitation signal is evolved from NB input speech signal. In the beginning, received NB speech undergoes LP analysis filter (inverse filter) which in turn brings forth NB excitation signal. As can be depicted from Fig. 5, NB excitation (e_{NB}) is then interpolated with factor of two followed by LPF to convert the NB excitation signal to WB sampling frequency rates. Interpolated and low pass filtered NB speech signal is then processed through FWR followed by HPF (having 3.4 kHz cut-off frequency). Analogous to other excitation generation techniques narrated in this paper, the role of FWR also extends towards typical expansion of the bandwidth of given signal. Further, WB excitations are produced by summing up interpolated NB excitation with HB portion of full wave rectified and high pass filtered excitation signal. Due to the rectification process, WB excitation signal has downward tilt at higher frequency and in order to compensate for the same, inverse filtering using whitening filter is performed on it (Nour-Eldin and Kabal 2008).

Sinusoidal transform coding approach is latest and predominant among all other techniques for generating WB residual signal from a given input NB recovered speech signal. As highlighted in Fig. 6, NB received input speech flows through LP analysis filter to produce baseband (NB) residual signal. This residual speech is then interpolated with factor two and subsequently passed from LPF having cut off frequency of 3.4 kHz. In order to produce HB excitation signal, NB interpolated and filtered residual signal is given to pitch detector and classifier block. To compute the degree of voicing classifier, approach mentioned in McAuley and Quatieri (1990), Uysal et al. (2005), and Cabaral and Oliveira (2005) is being explored for pitch estimation where spectral flatness measure (SFM) calculations can be computed on frame basis to separate out V/UV frames depending upon their energy in the given frame. Periodic and random excitations are then summed up and multiplied with gain factor. Eventually, WB excitations are produced by adding NB and HB excitations using over lap add method.

Significant demerit of STC lies in its estimation capabilities and accuracy of the pitch and degree of voicing; if violated, it does not preserve harmonic structure well between NB and HB also in the frame to frame basis. Formally, it is worth inspecting the addition of randomness in STC method, because, if unvoiced frames are synthesized only using periodic impulse train, tonal and buzz noise may occur with high intensity. Also, addition of noise to a periodic component is inadequate to eliminate periodicity of any given signal.

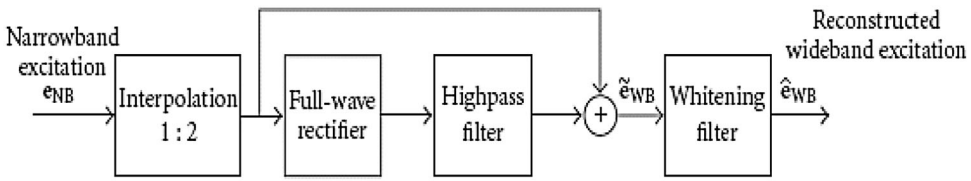

Fig. 5 Full wave rectification technique (Nour-Eldin and Kabal 2008)

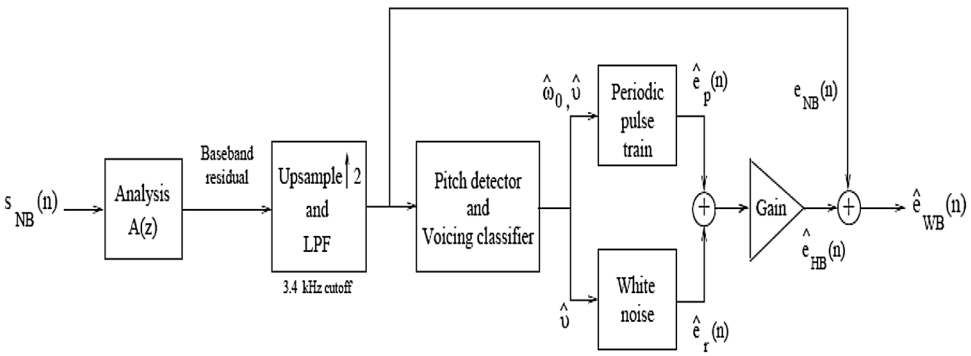

Fig. 6 Sinusoidal transform coding technique

Comparative studies and analyses between various extension of excitation methods reveal the fact that the results obtained in the cases of SF, ST, NLD and NM are quite comparable and quality of recovered WB speech (for all above cases) are also found superior in comparison with legacy GSM FR NB decoded speech. As far as the cases of SF and ST are concerned, because of incoherence in the phase of regenerated excitation, it may lead to degradation of periodic structure of voiced speech. However, it performs well in subjective analysis and listening tests because of the fact that discontinuity in the periodicity at the higher frequency may not be perceived by human auditory system. Though NLD is one of the competent candidates for WB excitation generation, it suffers from excessive aperiodic tones in HB and spectral aliasing. In NM, if the gaussian noise in HB is modulated by band pass filtered NB speech, the regenerated excitation preserves the harmonic structure without discontinuities. But simultaneously, if the gaussian noise is examined for modelling of HB excitation of voice frame, outcome is a noisy speech. Having distinguishable features, STC is one of the potential candidates among the others; however because of its implementation complexity in terms of requirement of accurate pitch and voicing estimators, its overall performance is subjective to computational accuracy of these estimators. An added as well as an inherent advantage offered by NLD and STC is its capability to reproduce missing frequency components in low band (50-300 Hz) frequencies (increased naturalness) without contributing much computational complexity into proposed system. Taking into account the factual aspect about harmonically poor structure of male speech in HB in contrast with female speech, subjective blind listening tests investigate better performance of male speech for various excitation regeneration methods over its counterpart. As stated earlier, in FWR method, rectification generates downward spectral tilt at higher frequencies of WB excitation and in order to compensate the same inverse filtering based on whitening filter can be applied but the accuracy of regeneration process relies in accurate designing of inverse filter and their cut off frequencies.

3.2 Implementation of MFCC based proposed ABE coder

As stated earlier, in MFCC based ABE coder the basic procedure of extracting and embedding HB features into NB bitstream of proposed GSM FR coder is analogous to LPC section. While comparing and contrasting both LPC and MFCC based ABE coders, it can be revealed that band splitting and recombining, watermark embedding and data hiding, extension of excitation sections and their implementations are identical which can also be witnessed from Fig. 7.

HB feature extraction based on MFCC This subsection aims at demonstrating the work of computing HB features with MFCC which can be embedded into NB bitstream and transmitted to receiver. One of the salient features of MFCC based parameterization lies in its usage of the human auditory system for speech parameterization. The vocal tract filter coefficients are estimated from the MFCC feature vectors. Figure 8 illustrates framework and process flow diagram to transform HB speech signal into equivalent representative MFCC feature vector on the frame wise basis.

This process commences by pre-emphasizing the speech which is used to flatten the spectrum before spectral analysis. Pre-emphasis block aims at compensating high frequency part of speech which was suppressed during speech production. HB speech is splitted into frames of 20 ms duration having total of 160 samples per frame. Hamming windowing is applied to reduce the discontinuity which may be introduced by framing process. Using fast Fourier transform (FFT) algorithm (k point), the magnitudes of spectral coefficients (abs(S(k)) of the speech frames are estimated. Further, task to be accomplished is to adept

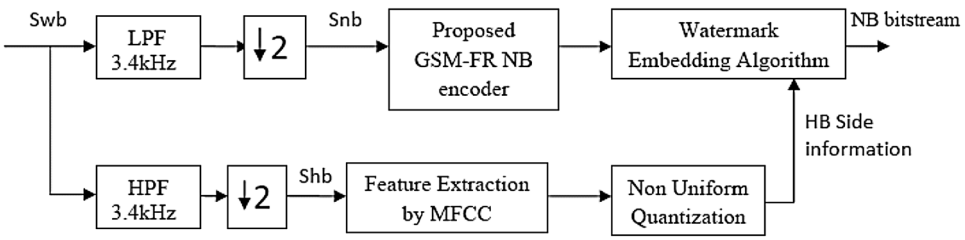

Fig. 7 Transmitter of proposed ABE coder based on MFCC features

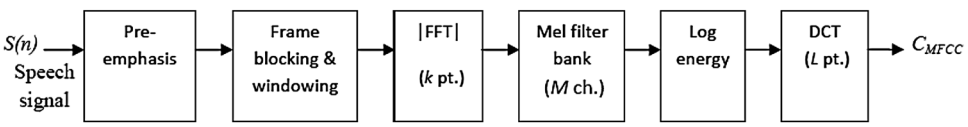

Fig. 8 MFCC parameterization of speech

frequency resolution to a perceptual Mel frequency scale. Filter bank consists of a set of BPF whose bandwidth and spacing are roughly equal to those of critical band and whose range of critical frequency covers the most important frequencies of speech perception. Principally, the filter banks are set of overlapping triangular BPF according to mel frequency scale. Centre frequencies of such filters are designed to be equally spaced below 1 kHz and logarithmically equally spaced above 1 kHz. Magnitude spectra (S(k)) of each frame serves as an input to mel filter bank and for each frame log spectral energy vector is produced as an output of filter bank analysis. Finally, mel filter-bank undergoes discrete cosine transform (DCT) to yield 13 MFCC coefficients (Nour-Eldin and Kabal 2008).

Furthermore, process of non-uniform quantization of MFCC coefficients allots bits as per their subjective importance as depicted in Table 3. Analogous to LPC based feature extraction, these coded 36 bits (spared bitrate of 1.8 kbps) are embedded and transmitted into bitstream of proposed GSM FR coder.

Wide band speech regeneration at receiver process is depicted in detail in Fig. 9 which demonstrates that at receiving end NB bitstream of 260 bits/frame undergoes watermark extraction algorithm which frame wise separates out the MFCC based HB features of 36 bits (1.8 kbps side information) from the input bitstream of 260 bits/frame. Next, NB bitstream is processed through proposed GSM FR decoder (legacy GSM FR NB decoder) to eventually reproduce NB speech after accomplishment of mathematical operations like interpolating it by the factor of two followed by passing it to LPF having cut-off frequency of 3.4 kHz. Subsequently, recovery of HB MFCC features from watermark extraction block is performed by conducting non-uniform dequantization. These frame wise extracted MFCC vectors are utilized to regenerate magnitude spectrum of the speech. Further, parallel to recovery of NB speech cited above, proposed GSM NB decoded speech (legacy decoded speech) also undergoes extension of excitation section followed by gain adjustment to yield WB residual signal. Several excitation regeneration techniques, as narrated in earlier subsection, are also adopted in this section of implementation for WB residual signal production. Only for such techniques in which LP analysis filter is not inbuilt, GSM NB decoded speech is first processed by LP analysis filter (using HB features) to generate NB residual signal before extension of such signals. LPC coefficients are derived from MFCC to LPC regeneration part (as will be described next) and supplied as it is to the analysis (only for above mentioned cases) and synthesis filter after interpolating it by two [for extension of envelop in line with Ramabadran and Jasiuk (2008)]. Role of synthesis filter is to produce WB version of speech from input WB residual signal, which is ultimately being processed through a HPF having 3.4 kHz cut-off frequency to yield HB portion of speech. Both HB and NB speech are finally integrated by summing them using over lap add method with necessary length equalization (if needed) to generate recovered WB speech.

HB synthesis or inversion of MFCCs to magnitude spectrum is a process which illustrates reconstruction a speech signal from a stream of MFCC vectors (frame wise) using a source-filter model of speech production that requires transformation of MFCC into the LPC coefficients of synthesis filter. For the same, estimation of magnitude spectrum is necessary which is achieved by inverting the MFCC feature vectors of HB speech. However, one of the

Table 3 Non uniform quantization and bit allocation of MFCC parameters representing HB features

Parameter	No. per frame	Resolution	Total bits/frame
MFCC coefficients	13	5,4,3,3,3,3,3,2,2,2,2,2,2	36
Total	36 bits		

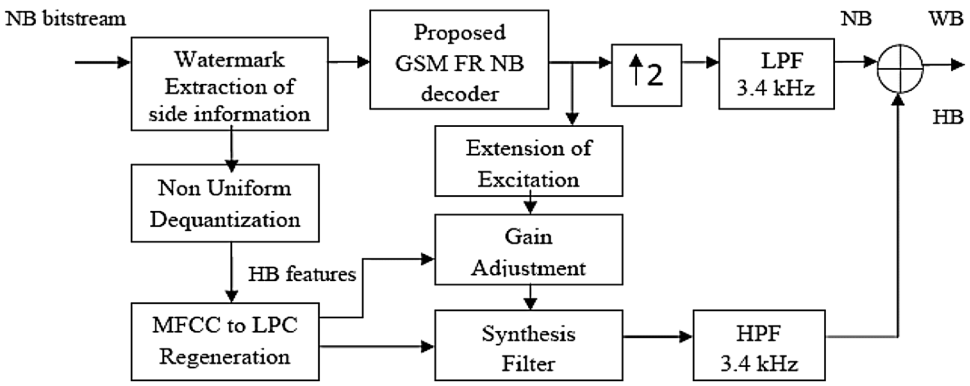

Fig. 9 Receiver section of proposed ABE coder based on MFCC

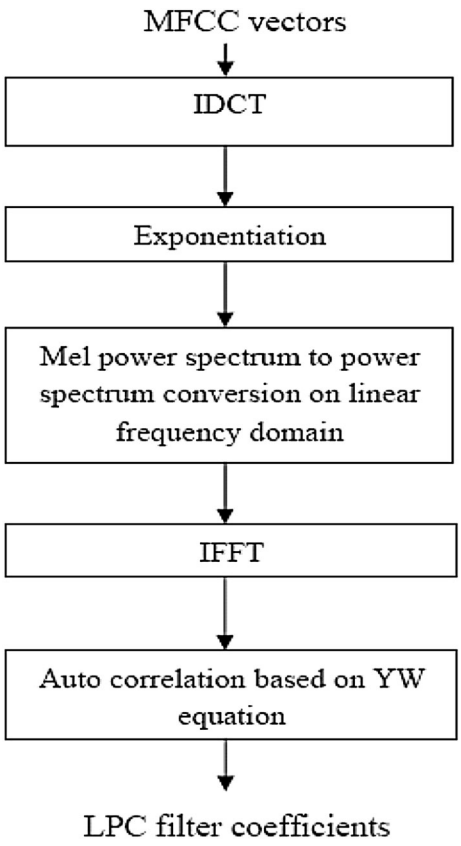

Fig. 10 MFCC parameters to LPC coefficients conversion

major limitations of MFCC based implementation is that the process of computing MFCC features described earlier is not completely invertible for some of the stages. As illustrated in Fig. 10, IDCT is applied for approximation as no direct transformation is possible due to less number of MFCC coefficients (which are 13) than the number of filter bank channels (which are 29). Samples of mel log power spectrum are uniformly spaced on the bandwidth of mel frequency domain by performing inversion of DCT followed by exponentiation process on obtained log mel power spectrum to yield mel power spectrum. Further, resultant power spectrum on the mel scale is converted into the linear frequency scale that eventually undergoes inverse FFT to generate the autocorrelation coefficients estimation. Finally, solving the Yule-Walker equation by autocorrelation caters the need of producing LPC coefficients of the synthesis filter (Nour-Eldin and Kabal 2008).

3.3 Comparative performance and consideration aspects between selection of LPC and MFCC techniques for HB feature extraction

In this research, framework pertaining to linear source filter model (LSFM) based approach is opted for implementation. Though MFCC to LPC transformation (which is not purely an invertible process) may exhibit more approximation process at receiving end, still, it remains the fact that MFCC representation contains more perceptual information compared to LPC parameterization and the same could be witnessed from obtained subjective and objective scores (as will be discussed next). Recent researches have also revealed the fact that MFCC parameterization results in better correlation between NB and HB (Nour-Eldin and Kabal 2008). The research reported in Nour-Eldin and Kabal (2008) drew an attention that MFCC have highest separability of speech class because of usage of DCT which may explore de-correlation of cepstral coefficients and also have second highest Mutual Information (MI) contents among different speech parameterizations. Such properties of MFCC are indeed evident for its wide applications in speech and speaker recognition. As stated in Nour-Eldin and Kabal (2008), HB certainty could be the ratio of MI to discrete HB entropy and this ratio offers better values for MFCC over line spectral frequencies (LSF) representing LPC coefficients. Increase in the ratio has been recorded with the increase in the MFCC dimensionalities which was not in the case of LSF. It should be brought to the notice that LPC to LSF conversion has not been implemented and hence not addressed in this research work.

Despite the fact about the benefits of MFCC over LPC based ABE implementation, LSF on the other side are quite preferable in speech coding applications because of their quantization error resilience performance and perceptual significance properties in terms of correlation between formant analysis and LSF pair computations (Nour-Eldin and Kabal 2008). An inherent merit in LSF is ease in conversion to equivalent LP coefficients and hence its wide applications in excitation estimation are evident; therefore LSF are the one among several potential candidates for ABE coder applications.

Thus, as discussed above, both LPC and MFCC based Proposed ABE coder stands out to be potential candidate for achieving backward compatibility with the use of modifications suggested in Proposed GSM FR NB coder, thereby yielding cost effective solution of implementing the same in stark contrast with legacy WB coders along with comparable recovered speech quality at receiving end.

4 Performance comparison and obtained results

In this paper, performance of proposed ABE coder based on feature extraction using LPC and MFCC techniques are evaluated using both objective and subjective analysis. Since each analysis is conducted on recovered WB speech yielded after implementing different extension of excitation techniques, for each of the case WB recovered speech can hence

be compared with proposed GSM FR (legacy GSM FR) decoder; also internally compared among themselves.

4.1 Subjective analysis

Subjective blind informal listening tests are conducted using mean opinion score (MOS) rating for four different clean wave files chosen from WB corpus cited in http://www.repository.voxforge1.org/downloads/SpeechCorpus/Trunk/Audio/Original/16kHz_16bit/. MOS analysis is conducted in quiet environment and with high quality headphones. For MOS analysis, twenty un-trained listeners are chosen to participate out of which ten listeners are men and other are women listeners. Each listener is randomly offered with total of 28 (four wave files of proposed GSM FR NB (legacy GSM FR) decoded speech and other 24 wave files of ABE WB recovered speech; between six extension of excitation methods each having four wave files) wave files for both of LPC and MFCC approaches. For both approaches, ratings given by all twenty listeners (for each individual case) are then averaged to produce final MOS ratings.

As observed from Tables 4 and 5 and Figs. 11 and 12, obtained results for MOS scores advocate the substantial improvement in performance of proposed ABE WB coder (for all wave files and for both approaches of LPC and MFCC) over its counterpart proposed GSM FR NB (legacy GSM FR) decoder. It is quite evident from the results obtained that MOS scores for all extension of excitation methods are quite comparable (in both of the LPC and MFCC approaches) but scores observed in STC and FWR techniques are found slightly better in comparison with the others. Results tested and highlighted in the bar graphs and tables are promising for both LPC and MFCC but it remains the fact that MFCC approach exhibits slightly better performance compared to LPC for few cases.

4.2 Objective analysis

In this work, it is convenient and sufficient to compute perceptual evaluation of speech quality (PESQ) scores for objective analysis of WB recovered speech quality as per P.862 standards (Rix et al. 2001). As depicted from Tables 6 and 7 and Figs. 13 and 14, PESQ scores recorded for different extension of excitation methods of proposed ABE coder (for both LPC and MFCC approaches) are found superior in comparison with PESQ score offered by proposed GSM FR NB (legacy GSM FR) decoder for all offered WB speech utterances. Analogous to subjective listening tests, FWR and STC outperforms other excitation regeneration techniques for both of the approaches with a few exceptions. Noteworthy feature of such examination is that both MOS and PESQ scores are quite comparable for all possible cases and significantly better compared to legacy NB coder. In stark contrast it is also visualized that for all cases, MFCC based proposed ABE coder performs slightly better in both analysis to its counterpart LPC approach on an average.

Table 4 MOS analysis of the proposed ABE coder based on LPC technique including different Excitation generation methods

Wave file	MOS Score for various approaches							
	No. of samples	Legacy GSM FR NB decoder	BWE using LPC technique with different extension of excitation methods					
			ST	SF	NM	NLD	FWR	STC
cc-17	64768	3.35	3.56	3.55	3.57	3.54	3.58	3.62
cc-21	52992	3.45	3.68	3.66	3.71	3.70	3.73	3.71
cc-24	106240	3.30	3.52	3.53	3.58	3.53	3.60	3.64
cc-27	125440	3.44	3.60	3.62	3.65	3.62	3.64	3.63

Table 5 MOS analysis of the proposed ABE coder based on MFCC technique including different excitation generation methods

Wave file	MOS Score for various approaches							
	No. of samples	Legacy GSM FR NB decoder	BWE using MFCC technique with different extension of excitation methods					
			ST	SF	NM	NLD	FWR	STC
cc-17	64768	3.35	3.61	3.60	3.62	3.59	3.64	3.72
cc-21	52992	3.45	3.67	3.66	3.69	3.71	3.59	3.68
cc-24	106240	3.30	3.54	3.52	3.59	3.58	3.65	3.67
cc-27	125440	3.44	3.62	3.65	3.64	3.59	3.63	3.71

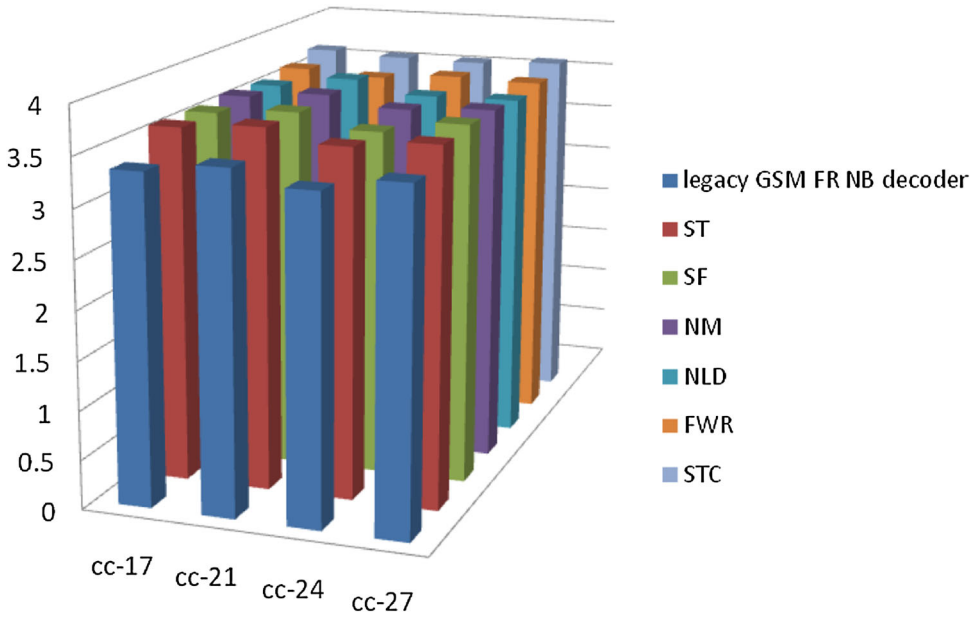

Fig. 11 Comparison chart for MOS score between all methods for LPC based proposed ABE coder

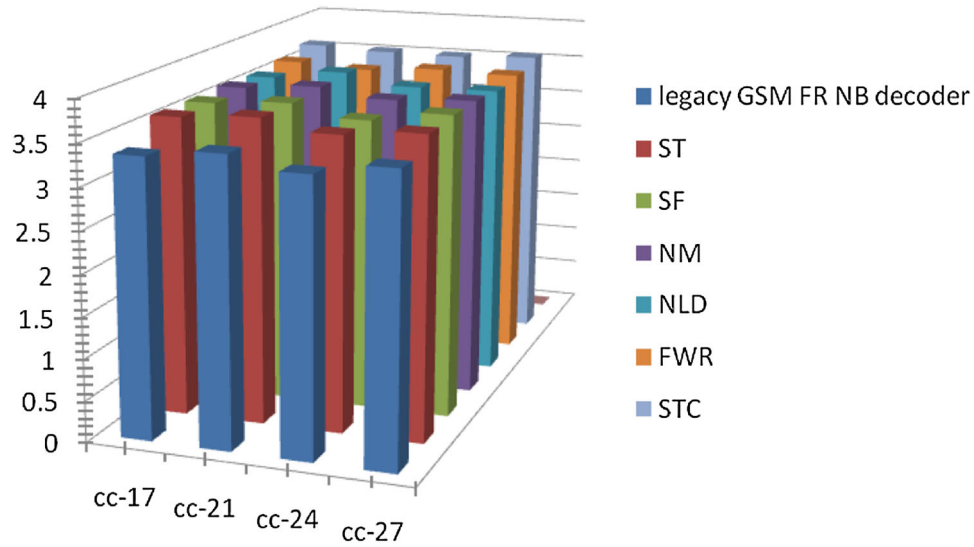

Fig. 12 Comparison chart for MOS score between all methods for MFCC based proposed ABE coder

Table 6 PESQ scores for proposed ABE coder based on LPC technique including different Excitation generation methods

Wave file	PESQ Score for various methods							
	No. of samples	Legacy GSM FR NB decoder	BWE using LPC technique with different extension of excitation methods					
			ST	SF	NM	NLD	FWR	STC
cc-17	64768	2.8089	3.1139	3.0468	3.1332	3.0879	3.1568	3.1759
cc-21	52992	3.0052	3.3092	3.2328	3.3160	3.2880	3.2896	3.3006
cc-24	106240	2.8240	3.0450	3.0873	3.0229	3.0742	3.1263	3.1527
cc-27	125440	2.8966	3.0910	3.0894	3.1481	3.1471	3.1379	3.1489

Table 7 PESQ scores for proposed ABE coder based on MFCC technique including different excitation generation methods

Wave file	PESQ score for various methods							
	No. of samples	Legacy GSM FR NB decoder	BWE using MFCC technique with different extension of excitation methods					
			ST	SF	NM	NLD	FWR	STC
cc-17	64768	2.8089	3.1066	3.0637	3.1176	3.0995	3.1665	3.1783
cc-21	52992	3.0052	3.3214	3.2507	3.3040	3.2837	3.3238	3.3192
cc-24	106240	2.8240	3.0722	3.0643	3.0959	3.0784	3.1062	3.1487
cc-27	125440	2.8966	3.0876	3.0754	3.1249	3.1523	3.1619	3.1728

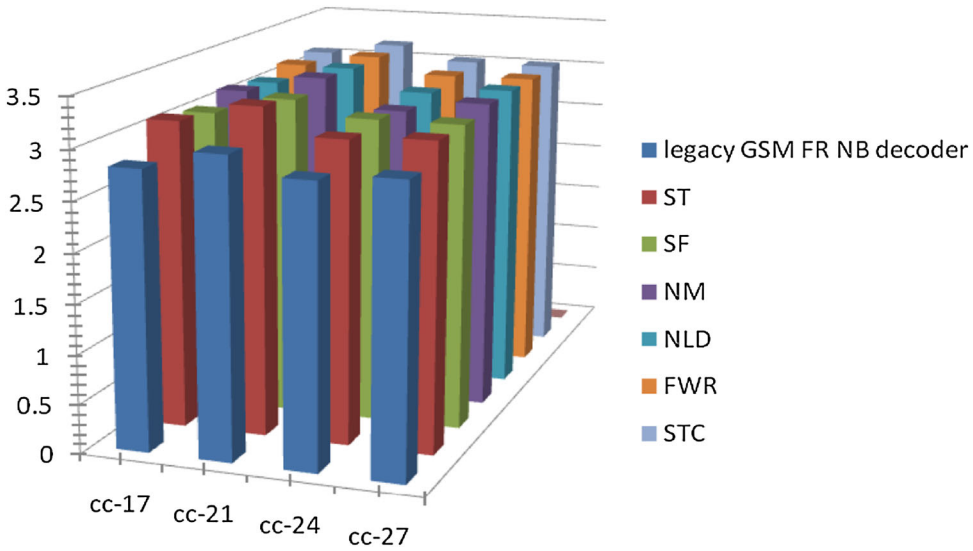

Fig. 13 Comparison chart for PESQ score between all methods for LPC based proposed ABE coder

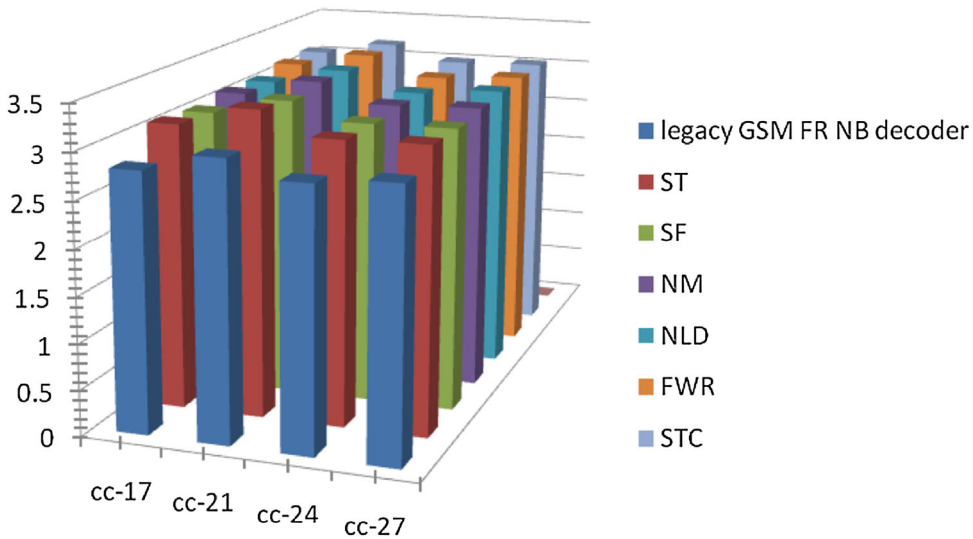

Fig. 14 Comparison chart for PESQ score between all methods for MFCC based proposed ABE coder

5 Concluding remarks

Acoustic bandwidth of existing NB transmission link is inherently limited which restricts overall intelligibility and naturalness of recovered speech and hence it sounds muffled and thin. Remedy to this bottleneck problem is inception of ABE which not only substantially improves quality of recovered speech (WB comparable speech) but also reasonably offers development of low computational complexity coders for portable devices.

This paper aims to provide a backward compatible solution to WB telephony over present scenario of NB transmission link by evolving proposed ABE coder with HB (side information) feature extraction, embedding, transmission and recovery. The proposed coder utilizes transmitted NB speech signal (proposed GSM FR NB coded speech) itself as a career of HB feature side information (required to conduct ABE) thus eliminating basic need of an additional channel. HB features (extension of envelop) representing side information are extracted by frameworks of LPC and MFCC based on Linear Source Filter Model. In this work, the payload of embedded data in terms of HB feature transmission is kept fix at 1.8 kbps (36 bits per frame) for both of the above frameworks. Further, joint source coding and data hiding/masking technique is adopted for embedding, transmitting and receiving HB side information over proposed GSM FR NB bitstream. These HB parameters, when decoded at receiver, are utilized to artificially reproduce WB speech and for the same several excitation regeneration techniques are devised, implemented and analyzed. Subjective informal blind listening tests based of MOS and objective evaluation of performance based on PESQ is computed and analyzed for the overall judgement of proposed ABE coder under various conditions of feature extraction as well as extension of excitations. Obtained MOS and PESQ scores for proposed ABE coder (for all cases) advocate significant improvement in comparison with the results evaluated for proposed GSM FR NB (legacy GSM FR) decoder. While comparing and contrasting different excitation regeneration methods for proposed ABE coder (for both LPC and MFCC frameworks), FWR and STC stand slightly superior among others as already demonstrated in bar graphs and tables. Still it remains the fact that for both analyses, all other excitation regeneration methods are also found comparable but values may vary subjectively depending upon the spoken utterances. Preliminary, listening tests based on MOS indicates that subjective speech quality (for both proposed ABE decoded and legacy NB decoded) is almost as good as objective scores suggest. Among both frameworks, MFCC parameterization is observed to be slightly better candidate over its opponent LPC which is quite evident from obtained results and graphs for both the analyses.

References

Bhatt, N., Gajjar, P., & Kosta, Y.(2012) Artificial bandwidth extension of speech & its applications in wireless communication systems: A review. In *Proceedings of IEEE international conference* on *communication systems and* network *technologies*, *Rajkot, India* (p. 563).

Bhatt, N., & Kosta, Y. (2011). Proposed modifications in ETSI GSM 06.10 full rate speech codec and its overall evaluation of performance using MATLAB. *International Journal of Speech Technology, 14*(3), 157.

Bhatt, N., Kosta, Y., & Tank, V. (2011). Proposed modifications in ETSI GSM 06.10 full rate speech coder for high rate data hiding and its objective evaluation of performance using simulink. In *International conference on communication systems and network technologies* (p. 27). Katra: IEEE Computer Society.

Cabaral, J., & Oliveira, L. (2005). Pitch-synchronous time-scaling for high-frequency excitation regeneration. In *INTERSPEECH* (p. 1513).

ETSI channel coding (GSM 05.03 version 8.9.0, release 1999,12, 2005-01).

ETSI digital cellular telecommunications system (phase 2+), full rate speech, transcoding, (GSM 06.10 version 8.2.0 Release, 10, 2005-06).

Fuemmeler, J., Hardie, R., & Gardner, W. (2001). Techniques for the regeneration of wideband speech from narrowband speech. *EURASIP Journal on Applied Signal Processing, 2001*(1), 266.

Geiser, B., & Vary, P. (2007). Backward compatible telephony in mobile networks: CELP watermarking & bandwidth extension. In *Proceedings of IEEE international conference on acoustics speech and signal processing* (*ICASSP*), *Toulouse.*

Jax, P., Geiser, B., Schandl, S., Taddei, H., & Vary, P. (2006). An embedded scalable wideband codec based on the GSM EFR codec. In *Proceedings of IEEE international conference on acoustics speech and signal processing* (*ICASSP*), *Toulouse.*

Jax, P., & Vary, P. (2003). On artificial bandwidth extension of telephone speech. *Journal of Signal Processing, 83*(8), 1707.

Jax, P., & Vary, P. (2006). Bandwidth extension of speech signals: A catalyst for the introduction of wideband speech coding? *IEEE Communications Magazine, 44*(5), 106.

McAuley, R. & Quatieri, T. (1990). Pitch estimation and voicing detection based on a sinusoidal speech model. In *IEEE transactions* on *acoustics*, *speech*, and *signal processing*, (*ICASSP*) (p. 249).

Nour-Eldin, A., & Kabal, P. (2008). Mel-frequency cepstral coefficient-based bandwidth extension of narrowband speech. In *INTERSPEECH* (pp. 53–56).

Ramabadran, T., & Jasiuk, M. (2008). Artificial bandwidth extension of narrow band speech signals via high band energy estimation. In *16th European signal processing conference* (*EUSIPCO*).

Rix, A. W., Beerends, J. G., Hollier, M. P., & Hekstra, A. P. (2001) Perceptual evaluation of speech quality (PESQ), and objective method for end-to-end speech quality assessment of narrowband telephone networks and speech codecs. *ITU-T Recommendation*, 862.

Shahbazi, A. (2010). Content dependent data hiding on GSM FR encoded speech. In *International conference on signal acquisition and processing*. Tehran: IEEE Computer Society.

Uysal, I., Sathyendra, H., & Harris, J. (2005). Bandwidth extension of telephone speech using frame-based excitation and robust features. In *13th European signal processing conference*, *Antalya.*

Vary, P., & Geiser, B. (2007). Steganographic wideband telephony using narrowband speech codecs. In *IEEE 41st Asilomar conference on signals, systems and computers* (*ACSSC*) (p. 1475).