Proc. of National Conference on Communications (NCC), 2007, Kanpur, India.

Signal Adaptive MBE Modeling for Low Bit Rate Speech Coding

Pushkar P. Patwardhan

Dept. of Electrical Engineering, IIT Bombay
Mumbai, India 400076
pushkar@ee.iitb.ac.in

Preeti Rao

Dept. of Electrical Engineering, IIT Bombay
Mumbai, India 400076
prao@ee.iitb.ac.in

Abstract

The Multiband Excitation (MBE) model has been found to be suitable for coding of narrowband speech at rates below 4 kbps due to its ability to produce natural sounding speech with a compact set of model parameters and its noise robustness. However the fixed frame analysis-synthesis framework of the MBE model is unsuitable for representing dynamic speech sounds such as stops and vocalic transitions. An alternative, the Signal Adaptive MBE model (SA-MBE), is proposed which adapts the time resolution of MBE analysis to the changing characteristics of the speech signal. Based on the transitory nature of speech either 10 or 20 ms analysis resolution is selected. The SA-MBE model is evaluated using subjective and objective methods and found to produce significantly better quality speech with improved intelligibility as compared to the fixed-frame MBE. Preliminary investigations suggest that the SA-MBE model parameters can be quantized to less than 3 kbps.

1. Introduction

The Multiband Excitation model [1] has been applied to low bit rate speech coding due to its ability to produce natural sounding speech from a compact set of model parameters and its inherent robustness towards noise. Based on sinusoidal modeling, an MBE codec has been standardized for satellite communication standard operated at 4.15 kbps [2]. In MBE modeling the short-time spectrum of the speech signal is represented using a fundamental frequency, a set of harmonic spectral amplitudes and a frequency dependent binary voicing function. The voicing function enables frequency domain multiband voicing decisions. The MBE modeled spectrum consists of bands of harmonic and noisy regions. The overall shape of the spectral envelope of the spectrum is specified by the envelope of spectral amplitudes of the harmonics.

Based on the assumption that the signal parameters are constant over the duration of the analysis window, the fundamental frequency, spectral amplitudes of the harmonics and voicing decisions are estimated once per 20 ms frame. While the quality of speech reconstructed from the model parameters is very good in most cases, perceptible degradations occur when the windowed signal does not satisfy the stationarity assumption. Aperiodicities in the glottal excitation and phonetic transitions within the analysis window are found to result in voicing artifacts in the reconstructed speech [3].

Various attempts have been made in the literature to improve the sinusoidal analysis-synthesis model. Early attempts included modification of analysis-synthesis methods for improved time varying amplitude modeling [4]. In [5], speech signal dependent biasing of the voicing decisions was proposed. In the context of sinusoidal models applied to speech signals, [6] presented an exponential harmonic sinusoidal model for representing time varying amplitudes of the harmonics. Recently, adaptive window size was used for analysis in a multiresolution sinusoidal model [7] and improved intelligibility was reported.

In this paper we propose a modification to the conventional MBE model with the aim to improve reconstructed speech quality with minimal increase in bit rate. An adaptive resolution analysis is proposed which serves to track rapid spectral changes while preserving spectral resolution in the steady or slowly varying regions of speech. Based on acoustic cues computed from the input waveform speech frames are classified as “transitory” or “non-transitory”. The non-transitory frames are modeled with a single 20 ms analysis, while the transitory speech frames are modeled using a pair of 10 ms analysis windows. Speech quality and intelligibility evaluation results are presented.

2. The MBE model

The conventional MBE model uses a fixed frame analysis-synthesis framework. The model parameters are estimated assuming a 20 ms stationarity interval. The MBE analysis algorithm estimates the model parameters (pitch and spectral amplitudes) by fitting an ideal harmonic spectrum to the actual spectrum [1]. For estimating the binary voicing decisions, the spectral mismatch between the actual and modeled spectrum is compared with a voicing threshold (which is dynamically updated) in each voicing band. A band in which the error exceeds the voicing threshold is marked as unvoiced. Of all the MBE model parameters the spectral amplitudes are the most demanding of the parameters on the bit allocation. We apply adaptive frequency warped LP modeling to model the spectral amplitudes with a 10^{th} order LP polynomial [8]. The LP polynomial is converted to an alternative LSF representation before quantization. The modeling of a vowel to nasal transition is shown in Figure 1. The temporal waveform of the transition is shown in Figure 1(a). It clearly shows that the shape of waveform changes rapidly across the pitch cycles although the pitch period remains the same. The spectra of the two 10 ms subframes are shown in Figure 1(b) and Figure 1 (c). The shape of spectra are observed to differ significantly. (The nasalization introduces a zero in low frequency region. This can be seen from the dip around 800 Hz in the spectrum of first subframe). Such time variations result in gross mismatch between the actual and modeled MBE spectra in 20 ms frame analysis and consequently a clearly audible voicing error.

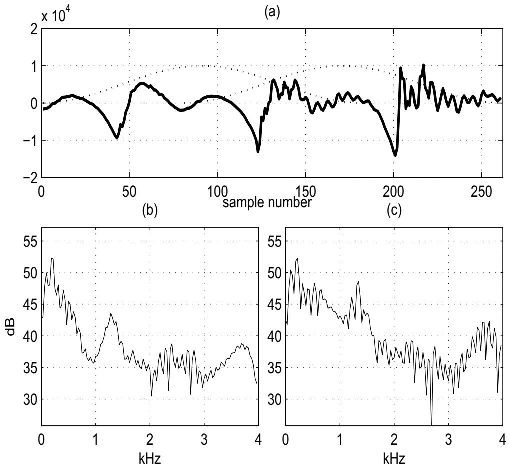

Figure 1: SA-MBE modeling of a transitional speech frame containing a vowel to nasal transition (transition "a-n-a", in phrase "a nice"). (a)Waveform. (b)Spectrum of first subframe. (c)Spectrum of second subframe.

MBE analysis assumes that the model parameters do not change over the analysis interval. However, nonstationarities are inherent in the speech signals due to aperiodicities in voicing and vocal tract dynamics. These are manifested as localized spectral and temporal transitions in the speech waveform. Due to the resulting mismatch with the assumed ideal harmonic spectrum, spurious unvoiced bands may be introduced in the reconstructed leading to significant degradation in the quality of modeled speech. Further, rapidly changing spectral and temporal features in case of speech waveform generated by stops and vocalic transitions cannot be reproduced by fixed frame MBE analysis. This results in a loss of intelligibility. To improve the quality of modeled speech in presence of nonstationarity, we present next a modification to the MBE model.

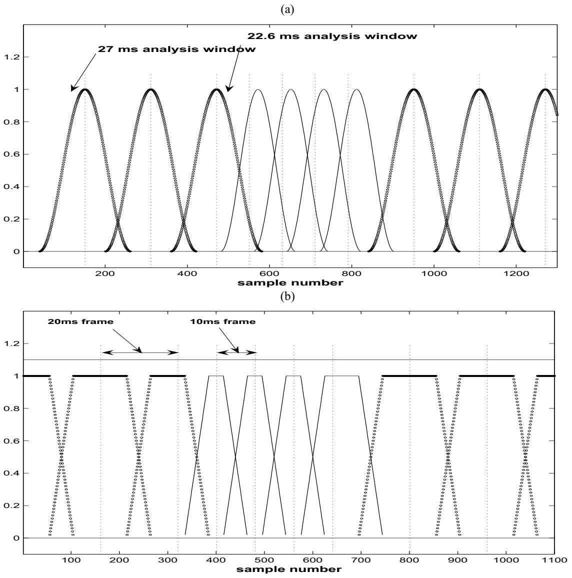

Figure 2: Placement of analysis and synthesis windows in SA-MBE modeling. (a)Analysis windows. (b)Synthesis windows.

3. The SA-MBE model

The algorithm is motivated from adaptive transform coding principles in audio compression schemes. The temporal resolution of the MBE analysis is adapted to the characteristics of the speech signal. The existing analysis framework is modified to adapt the time resolution of the analysis to the speech signal instead of using a single fixed frame size as in case of the conventional MBE. The new mode of analysis (with better time resolution) can be dynamically interleaved with the existing lower time resolution mode. The new functionality in the modified MBE model is the transitional frame detection mechanism. Apart from existing model parameters a flag indicating time resolution used for analysis is introduced. If the frame is identified as "transitory", instead of single set of model parameter, two sets of model parameters are computed with improved time resolution.

3.1. Detection of transients

The purpose of transient detection is to identify input speech frames containing transitions such as phone transitions, speech onsets and offsets, and possibly dynamic sounds such as plosives. The vocalic transitions such as those occurring in diphthongs, semivowels and vowel-nasal transitions are characterised by rapidly changing formant structure. The onset of voicing, end of voicing or stop consonants are characterised by abrupt energy change in temporal waveforms in the analysis frame. The algorithm used for detecting transient in this paper is based on computed temporal and spectral cues. These are the energy in low frequency region, ratio of spectral energies in low frequency region to high frequency region, ratio of energies in first half to the second half of the waveform in analysis window. On the other hand, vocalic transitions such as diphthongs and vowel-nasal transitions are detected by the Eucleadian distance between magnitude spectra of 10 ms subframes within the frame. The transient detection algorithm was tested on a large set of phonetically balanced sentences. A comparison with manually labeled transition frames showed that the algorithm was sufficiently accurate. It was found that only less than 10% of the automatically identified transient frames did not match with visually identified transient frames.

3.2. SA-MBE analysis and synthesis

The SA-MBE model allows the time resolution of MBE analysis to adapt to the signal characteristics. The SA-MBE modeling of a sequence of frames is illustrated in Figure 2(a). The figure shows analysis window placement for consecutive frames. The shorter analysis windows (of duration 22.6 ms) are interspersed with the longer analysis windows (which are 27 ms wide). The transition algorithm detects two frames in the region of 480 to 700 samples as "transitory". The analysis for these frames is carried out by a pair of short analysis windows. The short analysis windows have been designed based on considerations of duration and overlap. The short analysis window size is just adequate to estimate pitch periods of the order of 100 Hz, while low window allows pitch to be searched all the way till 70 Hz. Therefore in the SA-MBE model we choose not to estimate individual pitch periods for the subframes, but rather only once per frame. For frames analyzed with the short window, we estimate a pair of spectral amplitude vectors and pair of voicing decision vectors correpsonding to a pair of 10 ms subframes.

The MBE synthesis is similar to the conventional MBE synthesis, and is carried out for unvoiced and voiced speech independently [1]. If a 20 ms frame is to be synthesized, the procedure is identical to the conventional MBE procedure. If the frame is marked transitory, synthesis of 10 ms segments for each subframe is carried out separately to reconstruct the eventual 20 ms of the frame. Synthesis window positions for a sequence of frames containing a burst of transitory frames are shown in Figure 2(b). The add-overlap procedure must ensure that there are no distortions while synthesizing adjacent frames with differing transitory status. The addition of adjacent synthesis windows results in unity and no distortion occurs during the add-overlap synthesis. One must note that the shape of analysis and synthesis windows need not be same. The choice of analysis window in our case was Hanning since it is one of the most preferred window used for MBE parameter estimation [1]. The trapezoidal shape of synthesis window ensured "smooth" reconstruction of waveform in the overlap region between the adjacent frames. Other synthesis windows are possible, however one would need to remove the effect of window shape on the reconstructed waveform during the add-overlap synthesis.

		PESQ Score				
Sr no	Test item	MBE 20 ms	MBE 10 ms	SA-MBE 20/10 ms	% 10 ms frames	SA-MBE LSF intp.
1	"Why yell or worry over silly items"	3.39	3.75	3.61	21	3.22
2	"Beg that guard for one gallon of gas"	3.38	3.65	3.44	17	3.20
3	"Coconut cream pie makes a nice dessert"	3.22	3.56	3.35	20	3.15
4	"Nanny may know my name"	3.51	3.75	3.58	19	3.21
5	"His vicious father had seizures"	3.45	3.66	3.39	16	3.39

Table 1: Comparison of objectively evaluated qualities of MBE and SA-MBE modeled speech samples. MBE 20 ms: modeled with 20 ms analysis, MBE 10 ms: modeled with 10 ms analysis and SA-MBE: modeled with SA-MBE.

4. Experiment

A large number of sentences were used in the initial experimentation, based on their phonetic content and voice quality of speaker. The sentences were rich in particular class of phoneme (i.e. either voiced or unvoiced consonant, semivowel or diphthongs). Each sentence was modeled with the conventional MBE with fixed-frame 20 ms and fixed-frame 10 ms analysis, and the SA-MBE model (in which only transitional frames were switched to 10 ms mode). The quality of the modeled sentences was measured using the ITU P.862 PESQ algorithm [9]. Apart from objective evaluation, informal subjective tests were carried out to evaluate relative improvements in the quality of speech obtained with the SA-MBE model. The Table 1 illustrates the PESQ scores and the number of transient frames detected by the transient detection algorithm. The last column of the Table 1 shows PESQ scores for utterances with LSF interpolation based appproach applied for modeling of SA-MBE model parameters, which will be described in the following section. The Modified Thyme test [10] was conducted with six listeners to evaluate speech intelligibility.

5. Results and discussion

The results of objective evaluation of MBE and SA-MBE model are presented in Table 1. Clear improvement in quality of first four test items (which contain predominant vocalic transitions or voiced/unvoiced stops) with the 10 ms MBE analysis over the 20 ms MBE analysis is reflected by the increased PESQ scores. For the SA-MBE the actual percentage of frames switched to the 10 ms analysis mode is also shown in the second last column. We observe that by switching a relatively small fraction of frames to 10 ms analysis resolution the improvement in quality of modeled sentences is significantly improved towards that attained at overall 10 ms analysis resolution. The test item 5 contains mostly unvoiced fricatives and unvoiced sounds. It does not show any improvement with the SA-MBE modeling as 20 ms MBE mode is adequate. Informal subjective listening was entirely consistent with the objectively measured improvement. The improvement in intelligibility due to the SA-MBE is depicted in Table 2. The total % of rhyming words identified correctly during the intelligibility test is highest for those modeled with 10 ms MBE. The intelligibility with SA-MBE improves by roughly 6% on an average over the MBE. Improved intelligibility reflects the fact that the localized spectral / temporal transitions within the speech frames are reproduced better by the SA-MBE model than the conventional MBE model.

We see that rapidly varying spectral characteristics in the speech signal are captured more accurately with the higher time resolution provided by the SA-MBE model in the 10 ms mode. However a local increase in the framerate implies a corresponding local increase in the bit rate. This is undesirable for a fixed rate coder. It is known, from psychoacoustic studies mentioned in [11], that the human ear is less sensitive to degradation due to spectral distortion in case of transitory sounds. This implies that while high temporal resolution is important for the representation of transitory frames, spectral resolution can be reduced over that typically required for the representation of steady sounds. Based on this observation, we reduce bit-rate requirements by not explicitly retaining the LSF vector of the first subframe of the transitory frame. It was instead represented by linearly interpolating the LSF of the previous frame (or second subframe of previous frame in case it was a transitory frame) and the second subframe of the current frame. The interpolation factor was chosen from a 3-bit codebook based on the best match. The PESQ score of the SA-MBE utterance whose parameters were modeled using LSF interpolation technique described previously has been shown in the last column of Table 1. It is seen to be consistenly higher than 3.0. The PESQ score when compared with the PESQ score of utterance after SA-MBE analysis-synthesis only (shown in column 3 of Table 1) does not drop significantly. Informal subjective tests indicated that quality of speech quantized by this scheme did not significantly degrade the speech quality.

Based on some of the widely used bit allocation configurations for MBE model based speech coder in the literature [2], [12], our bit rate estimates suggest that the single pitch and a pair of voicing decision (one per subframe) can be quantized using 8 and 6 bits each. The gain per subframe can be quantized using 5 bits. The LSF vector can be quantized using 24 bits and a 3-bit interpolation coefficient. To indicate the 10 ms coding mode in SA-MBE, 1 bit per 20 ms frame is required, while single bit per frame is assigned to indicate the warping decision. This adds up to 53 bits per 20 ms for a transitory frame. For non-transitory speech frames, much finer quantization can be applied to model parameters to eventually obtain a fixed rate coder providing high quality speech at less than 3 kbps.

Algorithm	Frame rate (per sec)	%Intelligibility
MBE-20	50	82.9
MBE-10	100	91.1
SA-MBE	Adaptive	88.2

Table 2: Speech intelligibility as obtained by the MRT for the three configurations: MBE-20 ms, MBE-10 ms and SA-MBE.

6. References

[1] D. Griffin and J. Lim, “Multiband excitation vocoder,” *IEEE Trans. Acoust., Speech, Signal Processing*, vol. ASSP-36, no. 8, pp. 1223–1235, Aug 1988.

[2] J. Hardwick and J. Lim, “A 4.8 kbps multiband excitation coder,” in *Proc. IEEE Int. Conf. on Acoust, Speech and Signal Processing*, 1988, pp. 374–377.

[3] P. Rao and P. Patwardhan, “On representation of voice source aperiodicities in the MBE speech coding model,” in *Proc. of Workshop on Voice Quality: Functions, Analysis and Synthesis, Geneva*, 2003, pp. 27–29.

[4] M. Torres-Guijarro and F. Casajus-Quiros, “Improved analysis/synthesis methods for the multiband excitation codec,” in *Proc. IEEE Int. Conf. on Acoust, Speech and Signal Processing*, 1994, pp. 57–60.

[5] M. Zamrozik and J. Gowdy, “Modified multiband excitation model at 2400 bps,” in *Proc. IEEE Int. Conf. on Acoust, Speech and Signal Processing*, 1997, pp. 1603–1606.

[6] J. Jensen, S. Jensen, and E. Hansen, “Harmonic exponential sinusoidal modeling of transitional speech segments,” in *Proc. IEEE Int. Conf. on Acoust, Speech and Signal Processing*, 2000, pp. 1439–1442.

[7] K. Kim and I. Hwang, "A multiresolution sinusoidal modeling using adaptive analysis frame," in *Proc. of EUSIPCO*, 2004.

[8] P. Patwardhan and P. Rao, "Effect of voice quality on frequency-warped modeling of vowel spectra," *Speech Communication*, vol. 48, no. 8, pp. 1009–1023, Aug 2006.

[9] ITU-T, "Rec P.862, Perceptual evaluation of speech quality(PESQ) an objective assessment of narrow-band networks and speech codecs," ITU, 2002.

[10] S. Quackenbush, T. Barnwell, and M. Clements, *Objective measures of speech quality*. Prentice Hall, 1988, ch. 2.

[11] S. Wang and A. Gersho, "Phonetic segmentation for low rate speech coding," in Advances in Speech Coding, 1991.

[12] M. Nishiguchi, J. Matsumoto, R. Wakatsuki, and S. Ono, "Vector quantized MBE with simplited V/UV division at 3 kbps," in *Proc. IEEE Int. Conf. on Acoust, Speech and Signal Processing*, 1993, pp. 151–154.