On the Representation of Voice Source Aperiodicities in the MBE Speech Coding Model

Preeti Rao and Pushkar Patwardhan

Department of Electrical Engineering,
Indian Institute of Technology, Bombay, India 400076
Email: *prao@ee.iitb.ac.in, pushkar@ee.iitb.ac.in*

Abstract

We present an investigation of the representation of voice source aperiodicities in the Multi-Band Excitation (MBE) speech model for the compression of narrowband speech. The MBE model is a fixed-frame based analysis-synthesis algorithm which combines harmonic and stochastic components to reconstruct speech from estimated model parameters. Pitch cycle perturbations, such as jitter and shimmer, are not captured accurately in the framewise constant parameter estimates, thus impacting the reproduced voice quality. The actual dependence of MBE reconstructed voice quality on the voice pitch and the type of perturbation are explored through objective measurements and subjective listening with synthetic and natural speech.

1. Introduction

Over the years, there has been a move towards low rate coding of speech that preserves the naturalness of voice rather than just the intelligibility. Among the approaches particularly successful for low bit rate speech compression as well as speech synthesis are those from the family of hybrid sinusoidal models. Common to these models is the presence of sinusoidal components together with stochastic components. The Multi-Band Excitation (MBE) model [1] is an example of a harmonic-stochastic speech model, widely considered to be suitable for the coding of narrowband speech at bit rates below 4 kbps. The MBE model represents the spectrum of an input speech frame of fixed duration (typically 20ms) as the combination of distinct frequency bands some containing pure harmonics of the fundamental frequency (“voiced” bands), and others spectrally shaped random noise (“unvoiced” bands). Thus the MBE model generalises the traditional single voiced/unvoiced decision per frame into a set of decisions, each representing the voicing state within a particular band. The chief motive for this “mixed” representation, as depicted by Figure 1, is to improve the reproduction of mixed-voice speech segments such as voiced fricatives, as well as the representation of speech in acoustic background noise.

While the spectra of frames of steady vowels are expected to be adequately represented by the sum of harmonics, the harmonic structure is often altered due to the presence of glottal source perturbations typical in natural speech. These include pitch period perturbations (jitter), amplitude perturbations (shimmer), and the presence of aspiration noise in portions of the glottal vibration cycle. While the perceived naturalness of normal voices owes to an extent to the small period-to-period variations of glottal source waveform, emotionally uttered as well as disordered voices are typically characterized by moderate to large perturbations of the glottal vibration waveform. In general auditory perceptual terms, additive noise is associated with breathiness, with moderate amounts of jitter and shimmer being associated with roughness [2]. In general physical terms, noise is associated with turbulent flow at the glottis and perturbation (jitter or shimmer) with vocal-fold instability. A perceptually accurate reproduction of the effects of such glottal source perturbations is highly desirable for natural sounding speech.

The goal of the present work is to explore the representation of voiced speech characterized by glottal source perturbations by the MBE speech coding model. As such the only attribute available to capture any voice source aperiodicities within the analysis window of a frame is by the distribution of voiced and unvoiced bands across the speech spectrum. In this context, the observations of Fujimura [3] on the role of mixed excitation in a channel vocoder are relevant. Via subjective listening experiments, Fujimura showed that a crude approximation of aperiodicity (of various causes as observed in natural speech) can be made by distributing patches of random noise signals in the time-frequency space of the speech signal. The devoicing of higher frequency vocoder channels was achieved by using random noise as the source excitation in these channels. Griffin and Lim [1] base their justification of the advantage of multiband excitation over single band excitation on the observations of Fujimura, in addition to their own observations on speech spectra corrupted by random, additive noise. On the other hand, Makhoul et al [4] note that spectral devoicing due to vocal cord vibration irregularities is strictly speaking an artifact of the spectral estimation process, and therefore it may not be appropriate to use a noise source to synthesize devoiced spectral regions due to the various forms of voice aperiodicities. They use a simple voicing cut-off scheme in their “mixed source” model intending only to represent accurately the turbulent source that is likely to occupy the high frequency region of the spectrum.

In view of the above discussion, we examine the MBE speech coding model representation of narrowband voiced speech characterized by glottal source perturbations. The next section reviews the MBE model with emphasis on those features relevant to the present study. In the following sections, the influence of voice source aperiodicities on model parameters and on the reconstructed speech are examined by way of synthetic vowel sounds and natural speech samples selected from the VOQUAL’03 database [5].

2. MBE Speech Coding Algorithm

The speech model underlying the MBE speech coding algorithm is a special case of the general harmonic-stochastic model. The MBE speech model disregards harmonic phases and assigns frequency-dependent voiced/unvoiced decisions, with the decisions being associated with bands each containing a fixed number of harmonics (typically 3) [1],[6]. The modeling is implicitly tied to the analysis and synthesis procedures. In the analysis step, the input speech signal is divided into successive, overlapping frames at a frame rate of 50 frames/sec. Within each frame, the signal is assumed to satisfy a constant-amplitude, constant-frequency sinusoidal model. The model parameters are estimated by an analysis-by-synthesis method by minimizing the distance between the original and an assumed synthetic speech spectrum. The error distance is first minimized over the fundamental frequency and spectral amplitudes assuming all voiced speech. Once these parameters are estimated, voicing decisions are made for each band of three harmonics based on the closeness of fit between the original and synthesized spectrum for each group of harmonics. A predetermined frequency-dependent threshold is applied to this normalized error to obtain a voicing decision for each band. The accuracy of the pitch estimate is crucial since both spectral amplitudes and voicing decisions are based on the match of a synthetic spectrum of the estimated pitch, with the input spectrum. Gross pitch errors due to the selection of pitch period multiples are minimized by a dynamic pitch tracking method which favours lower submultiples of pitch period.

At the decoder, unvoiced bands are reconstructed by spectrally shaped random noise. Voiced bands are synthesized as the sum of harmonics to be combined across frames by either overlap-add or interpolation, depending on the extent of pitch difference between frames. At low bit rates, phases of the harmonics are not transmitted but are regenerated at the decoder so that the phase evolves smoothly across frames. The regenerated phase has an important influence on the voice quality. A controlled amount of frame-to-frame phase randomisation of harmonics is introduced whenever the frame contains at least one unvoiced band to improve naturalness [6].

3. Effect of Voice Source Perturbations

In the production of voiced speech, source aperiodicities (i.e. variation in glottal excitation waveform from period to period) can arise either from additive random noise or from modulation aperiodicities such as jitter, shimmer and prosodic pitch variations. All normal voices contain vocal jitter, generally less than 1%. In disordered voices , it is much higher and is correlated with perceived roughness. Vocal shimmer is the cycle-to-cycle variation in amplitude, which in normal voices is less than 7% of the mean amplitude. Additive random noise or aspiration noise contributes to the sensation of breathiness. It has a nearly flat spectrum but is perceptually significant in the higher frequency region where it is not masked by strong speech harmonics. Both roughness and breathiness, as well as strong variations of pitch due to intonation, are important components of emotionally uttered speech.

As discussed in Section 2, MBE speech analysis is primarily a frequency-domain harmonic spectrum matching procedure. That is, the parameters of pitch, spectral envelope and multiband voicing decisions in each frame are estimated based on the error between the spectrum of the windowed input speech and the spectrum of an assumed purely periodic signal with the same window applied. When the window contains several pitch periods, as is the case for high-pitched voices, the presence of source aperiodicities perturbs the regular harmonic structure of voiced sounds. Modulation aperiodicities lead to the broadening of harmonic main lobe bandwidths while aspiration noise contributes to noise between harmonics. Jitter leads to a smearing of the harmonic structure that is more prominent at higher frequencies due to the greater frequency deviations for a given percentage jitter.

The disturbance of the harmonic structure, as discussed above, leads to an increase in harmonic spectrum matching error in the corresponding MBE model frequency bands. Whenever the error exceeds the voicing threshold for the band, the band is reconstructed by MBE synthesis as spectrally shaped random noise. So we see that both types of voice source aperiodicities lead to the synthesis of the sound as a mixture of frequency bands containing harmonics only and bands containing purely random noise.

The bandwise voicing pattern varies in time (i.e. across frames). When the pitch change from one frame to the next one is within 10% of the estimated pitch, the frames are combined through linear interpolation of the amplitudes and frequencies of the corresponding voiced bands. When the extent of pitch change is higher or an unvoiced band is involved, overlap-add is used to achieve the final synthesized speech.

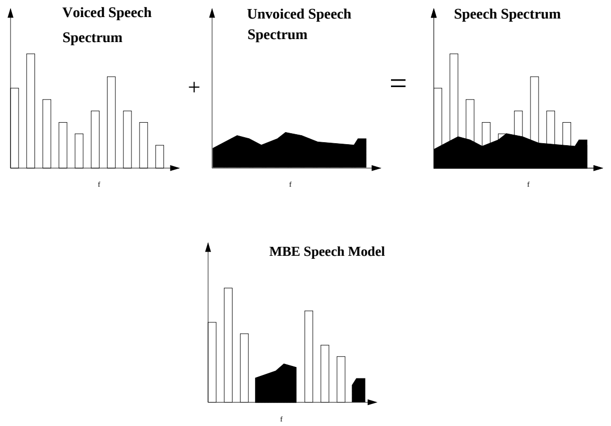


Figure 1: MBE representation of the speech spectrum [7]

4. Experimental Results

4.1. Synthetic vowels

The vowel ɑ (as in the word “harm”) was synthesized at three different fundamental frequencies by filtering an LF model glottal pulse train [8] by vocal tract parameters derived from a natural sound uttered by a low pitched male speaker. Controlled amounts of jitter and shimmer were each introduced into the glottal pulse train to generate a test set of steady voiced stimuli using the algorithm of [9]. Two values each of percentage jitter and shimmer, one below the threshold for abnormality and the other above, were simulated. The duration of each sample was 600ms, and the two ends of the segment were tapered for better listening. The test sounds were processed by MBE analysis and synthesis programs to obtain the corresponding “modeled” sounds.

It was observed that even at the higher extents of pitch perturbation, the MBE estimated pitch lay within the range of that of the reference sound. Pitch octave errors, likely to occur in the presence of pitch cycle perturbations, were effectively suppressed by the MBE analysis pitch tracker. Table 1 shows the measured values of jitter and shimmer before and after MBE modeling of the test sounds. The percentage jitter/shimmer were computed as the mean absolute jitter/shimmer in percentage across the steady portion of the sound [10]. Also shown in Table 1, are values representing the percentage of the time-frequency space of each sound that is devoiced as determined by the MBE band voicing decisions. These regions are reconstructed as spectrally shaped random noise by the MBE synthesis program.

	Jitter			Shimmer		
	% jitter			% shimmer		
Pitch (Hz)	ref.	mod.	% devoicing	ref.	mod.	% devoicing
80	0.6	0.3	0	6.0	5.2	0
	2.7	1.5	0	10.0	9.4	0
160	0.6	0.1	0	3.9	1.0	0
	2.9	1.4	10	9.4	2.6	0
250	0.6	0.05	0	3.7	0.5	0
	2.9	0.7	22	9.6	1.0	0

Table 1: Objective measurement results for pitch cycle perturbed sounds before (ref.) and after (mod.) MBE modeling.

We observe from Table 1 that the % jitter retained in the modeled sound with respect to that of the reference sound decreases with increasing fundamental frequency. The same is true for % shimmer. The extent of devoicing increases with increasing perturbation in the case of jitter, but is not affected at all in the case of the shimmered samples. This implies that, the harmonic broadening due to shimmer is not sufficiently high to cause the spectral error to cross the voicing threshold. Informal listening revealed that the perceptual attributes of the MBE modeled sounds were in correspondence with the objective measurements of Table 1. That is, the roughness sensation induced by the aperiodicities is nearly retained in the low-pitched vowel but is diminished as the fundamental frequency increases. The unvoiced regions introduced in the modeled sounds are mainly in the higher frequency region, and they do not contribute much to the perceived roughness. Rather they were found to give rise to a perceptually “unfused” noisiness.

4.2. Natural speech

VOQUAL’03 database samples of utterances characterized by voice aperiodicties (such as harsh, rough and breathy voices) were processed by MBE analysis-synthesis. An audio demo can be found at [11]. As expected from the results on synthetic vowels, voice quality due to glottal cycle aperiodicities is preserved for low pitched voices. It was found that breathy voices are reconstructed with no perceptible degradation at all pitches due to the appropriate introduction of unvoiced bands in the MBE modeled sounds.

To investigate the MBE modeling of high-pitched voices, several samples from the VOQUAL'03 database were studied closely, one of which is discussed next as a representative case. Figure 2(a) shows the waveform and spectrogram of a Japanese utterance from VOQUAL'03 database (downsampled to 8 kHz). The pitch range is 170 to 400 Hz. The sample is characterized by roughness coupled with breathiness during the first syllable (0-0.3 sec)and portions of the second syllable (0.3-0.6 sec). The spectrogram was computed with a window length of 28 ms and a hop size of 2.5 ms. Salient features of this "harsh" speech sample are: 1. prominent prosodic pitch variations in the first and third syllables; 2. modulation aperiodicities coupled with aspiration noise during the first and second syllables. Figure 2(b) shows the spectrogram of the same utterance after MBE analysis and synthesis. Perceptually it is found that the roughness is retained while the breathiness has been lost to an extent in the first syllable.

The spectrogram of Figure 2(b) provides interesting insights into the MBE modeling of the pitch cycle variations. We observe a "net" like time-frequency structure during portions of the first syllable. This is due to the oscillation of band voicing decisions, between voiced and unvoiced, across frames in time. The overlap-add synthesis then leads to the simultaneous presence of strong harmonics and noise in the low frequency bands resulting in a sensation of roughness. We also see the replacement of aspiration noise in the higher frequency region by harmonics. This reduces the breathiness of the voice perceptibly. In the second syllable, the roughness is replaced by breathiness due to the replacement of weak low harmonics by noise. The high frequency bands show some devoicing in parts (contributing to breathiness) but otherwise show clear harmonics. Finally, the rapid pitch increases during the first and third syllables cause discontinuities in the pitch tracks of the framewise constant estimates in MBE synthesis. In the higher frequency region of the third syllable, we note the appearance of patches of noise which lend a breathy quality. This is attributed to the increased broadening of the higher harmonic mainlobe bandwidths due to the pitch variation. This phenomenon (also noted in the context of wideband synthesis by [12]) leads to the interpretation of harmonic components as noise in MBE analysis which is biased toward the detection of constant amplitude-constant frequency harmonics within the analysis window.

5. Conclusions

The above experimental results indicate that at low fundamental frequency, the pitch cycle perturbations as reflected in the objective measurements of jitter and shimmer are reproduced, to a great extent, in MBE modeled speech. This explains why the perceived voice quality is preserved at low pitches. At higher values of fundamental frequency, the perturbations are not captured to the same extent in the modeled sounds due to the short duration of the pitch cycle relative to the analysis window duration.

To summarise our observations, for voiced speech signals exhibiting source aperiodicities, the MBE parameter estimation algorithm cannot track the temporal fluctuations in the pitch cycle accurately. At low fundamental frequencies (long pitch periods) there are only 2 or so periods per frame implying that MBE analysis is able to track pitch and amplitude variations to a large extent. At high fundamental frequencies, the frame-based analysis algorithm averages the pitch cycle variations over the window length. This leads to the suppression of the temporal fluctuations. However, in the analysis, the harmonic spectrum matching error goes up, leading to the possibility of interspersed unvoiced bands. This is due to the inability of the frame based analysis algorithm to differentiate between noise and modulation aperiodicity. (That this is true more generally is seen in [13]) Thus, in the case of high fundamental frequencies, all types of glottal waveform perturbation (jitter, shimmer, aspiration noise) give rise to regions of devoicing in the time-frequency plane. When the devoicing occurs in the low frequency region and coexists with low frequency harmonics, a sensation of roughness is synthesized. Thus subjective listening reveals that the mixed-voiced representation of the MBE model does to a limited extent serve to simulate the perceptual effects of roughness and, to a much greater extent, breathiness.

The fixed frame analysis-synthesis of the MBE speech coding algorithm is the chief cause of the perceived voice quality degradation. Reducing the frame size or using a pitch adaptive framework would improve performance for high pitched voices. But this is typically not an available option in fixed-rate speech coding.

6. References

[1] Griffin D.W.and Lim J.S, "Multiband excitation vocoder," *IEEE Trans. Acoustics, Speech and Signal Processing*, vol. 36, no. 8, pp. 1223–1235, Aug 1988.

[2] Hillenbrand J., "Perception of aperiodicities in synthetically generated voices," *J. Acoust. Soc. Am.*, vol. 83, no. 6, pp. 2360–2371, June 1988.

[3] Fijumura O., "An approximation to voice aperiodicity," *IEEE Trans. on Audio and Electroacoustics*, vol. AU-16, no. 1, pp. 68–71, March 1968.

[4] Makhoul J., Vishwanathan, Schwartz R., and Huggins A.W.F., "A mixed-source model for speech compression and synthesis," *J. Acoust. Soc. Am.*, vol. 64, no. 6, pp. 1577–1581, Dec 1978.

[5] VOQUAL'03 Workshop, "Voice material for VOQUAL'03,"

http://www.limsi.fr/VOQUAL/voicematerial.html, Aug 2003.

[6] Kondoz A.M., *Digital Speech: Coding for Low Bit Rate Communication Systems*, Chapter 8, John Wiley, New York, 1991.

[7] “INMARSAT M VOICE CODEC,” Digital Voice Systems Inc., Aug 1991.

[8] Veldhuis R., “A computationally efficient alternative for the Liljencrants-Fant model and its perceptual evaluation,” *J. Acoust. Soc. Am.*, vol. 103, no. 1, pp. 566–571, Jan 1998.

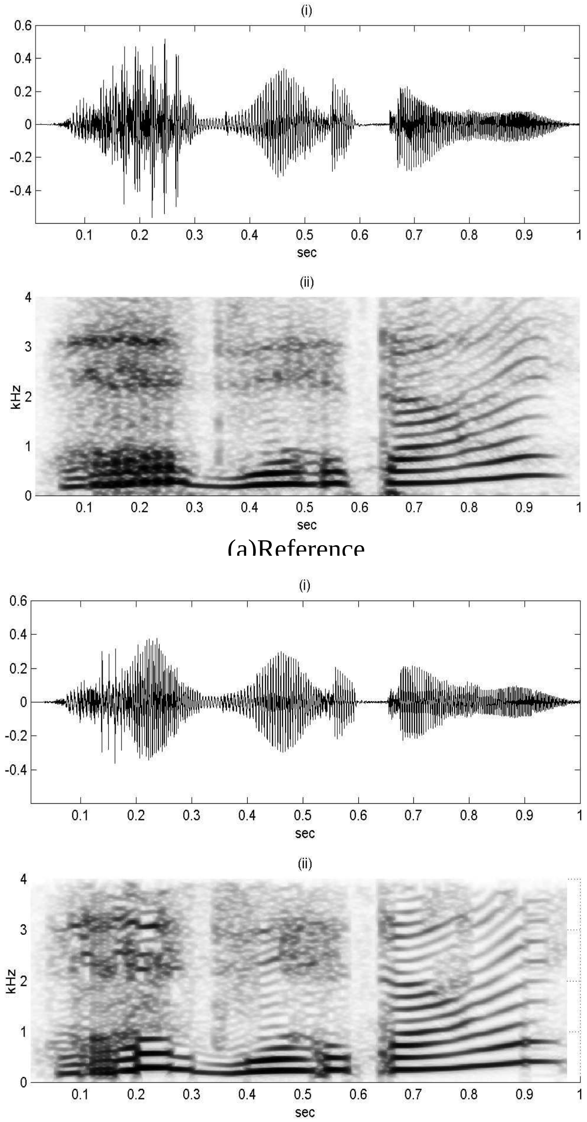
[9] Lalwani A.N. and Childers D.G., “Modeling vocal disorders via formant synthesis,” in *Proc. Int’l. Conf. on Acoust. Speech. and Signal Processing*, 1991, pp. 505–508.

[10] Boersma P. and Weenink D., “Praat: Doing phonetics by computer, verion 4.0.51,” http://www.praat.org/, 2003.

[11] Rao P. and Patwardhan P., “VQW-03 demo. samples,” http://www.ee.iitb.ac.in/˜prao/speechcoding/ voqual03/index.html, May 2003.

[12] Dutoit T. and Gosselin B., “On the use of a hybrid harmonic/stochastic model for tts synthesis by concatenation,” *Speech Communication,*, vol. 19, pp. 119–143, 1996.

[13] D’Alessandro C., Darsinos V, and Yegnanarayana B., “Effectiveness of a periodic and aperiodic decomposion method for analysis of voice sources,” *IEEE Trans. on Speech and Audio Processing*, vol. 6, no. 1, pp. 12–23, Jan 1998.



(a)Reference

(b)Modeled

Figure 2: Waveform and spectrogram of Japanese utterance [he:jie:itai] (Harsh-391) from VOQUAL’03 database. (a) Reference speech (b) MBE modeled speech