MODIFIED MULTIBAND EXCITATION MODEL AT 2400 BPS

Michele Jamrozik and John Gowdy

Electrical & Computer Engineering
Clemson University, Clemson, SC 2963-0915 USA
mjamroz@eng.clemson.edu & john.gowdy@ces.clemson.edu

ABSTRACT

This paper presents the Modified Multiband Excitation Model used for speech coding. In many MBE model coders, speech quality is degraded when incorrect voicing decisions are made, particularly for high-pitched female speakers. The MMBE addresses this issue with a modified voiced/unvoiced decision algorithm and a more robust pitch estimate. The listening quality of speech produced using the MMBE model is superior to the FS-1016 CELP coder and is at least comparable with the new 2400 bps MELP coder chosen as the new 2400 bps Federal Standard.

1. INTRODUCTION

In the past, vocoders have been capable of producing intelligible but not high quality speech at bit rates of 4800 bps and below. The poor quality of the synthetic speech in these vocoders can, in part, be attributed to the fundamental limitations of the speech models and, in part, to the inaccurate estimation of the speech model parameters.[1] The Multiband Excitation (MBE) model first proposed by Daniel Griffin at MIT, has addressed some of these limitations and is a valid candidate for low bit rate applications.

The model presented in this paper, denoted the Modified Multiband Excitation (MMBE) model, is an outgrowth of some existing MBE models [1, 2, 4, 5] with some added enhancements. The MMBE model is very similar to [1] in that an analysis by synthesis approach is taken to develop the parameter set and multiple voiced/unvoiced decisions are made for each speech frame. In general, MBE models represent a synthetic speech signal $s(n)$ as the response of a linear filter $h(n)$ to some excitation signal $e(n)$ such that

$$S(\omega) = H(\omega)E(\omega) \qquad (1)$$

where $H(\omega)$ and $E(\omega)$ are the Fourier transforms of $h(n)$ and $e(n)$, respectively.

The major difference between the MBE models and traditional vocoders is the way the excitation signal for each speech frame is determined. In previous vocoders, an entire speech frame was declared either voiced or unvoiced. In contrast, the MBE models divide the original excitation spectrum into a given number non-overlapping frequency bands and a V/UV decision is made for each of these bands. An appropriate excitation sequence is then generated for each of these frequency bands depending on the voicing decisions.

The synthetic speech resulting from multiple voicing decisions better represents speech frames having both voiced and unvoiced components. Therefore, speech quality improves and the "buzziness" caused by replacing noise-like energy in the original spectrum with periodic energy in the synthetic spectrum is reduced. [1] Figure 1 shows an example of this for the voiced fricative \z\ which contains both voiced and unvoiced components.

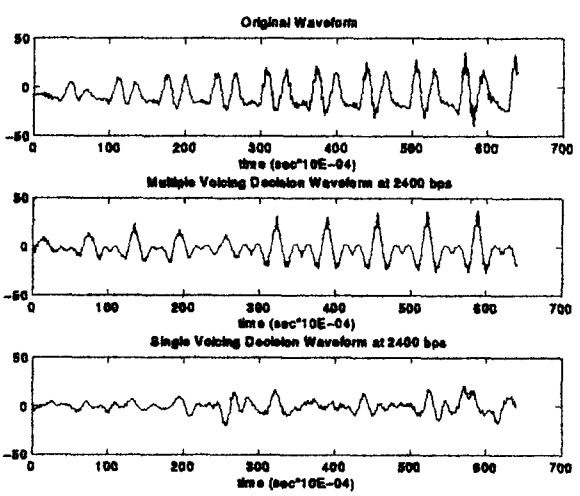

Figure 1: Comparing Multiple to Single Voicing Decisions for Voiced Fricative \z\ at 2400 bps

2. MMBE

The enhancements exclusive to the MMBE model include a modified pitch tracking algorithm and a more robust voiced/unvoiced decision making scheme. Other features of the MMBE model have been adopted from the various versions of the MBE. These features include limiting the number of voiced/unvoiced decisions per speech frame to twelve as in [4] and using DCT coefficients to represent the spectral envelope as in [5]. Figure 2 is a block diagram of the MMBE.

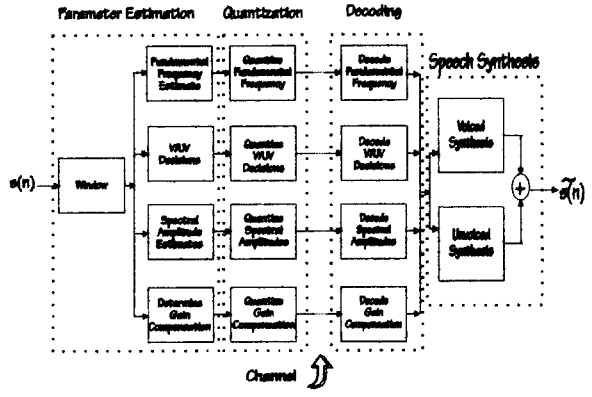

Figure 2: Block Diagram of the Modified Multiband Excitation Model

3. ESTIMATING THE SPEECH MODEL PARAMETERS

The MMBE parameter estimates are based on how closely the synthetic speech matches the original speech signal. Therefore, the quality of speech generated using the MMBE model is highly sensitive to the accuracy of the estimates of the speech model parameters. The MMBE focuses on generating more accurate parameter estimates than previous MBE models. The MMBE model parameter set includes the fundamental frequency (pitch), voiced/unvoiced decisions, and the spectral envelope for each speech frame.

3.1. Pitch Estimation

This section describes the process involved in determining the pitch value $\hat{P}_0$ of the current speech frame. The pitch value is related to the fundamental frequency $\hat{\omega}_0$ in the following manner:

$$\hat{\omega}_0 = \frac{2\pi}{\hat{P}_0} \quad . \tag{2}$$

The units of $\hat{w}_0$ are radians/sample.

The proposed pitch algorithm reduces pitch estimate errors and is composed of several steps. First, error values ϵ for all possible pitch values are determined. Once these error estimates are made, both forward and backward pitch tracking algorithms are employed. These pitch tracking algorithms limit the range of acceptable pitch values to help avoid sharp discontinuities in the pitch periods of successive frames.

In most MBE models, voicing decisions are based on how closely the original spectrum matches the synthetic spectrum generated using the current pitch estimate. Therefore, all frames, whether voiced or unvoiced, have a pitch estimate associated with them. Tracking a pitch value for an unvoiced frame can lead to invalid pitch estimates for voiced frames as can be seen in the top portion of Figure 3. To avoid this, the backward pitch tracking scheme for the MMBE has been modified such that

$$P_{-1} = \begin{cases} \hat{P}_0 & \text{if } sum(\vec{v}) \geq \frac{length(\vec{v})}{2} \\ P_{-1} & \text{otherwise} \end{cases} \tag{3}$$

where $\vec{v}$ is the vector containing all the voicing decisions for the current frame and P_{-1} is the pitch value of the previous frame. This means, if the majority of a speech frame's voicing decisions are declared unvoiced, the current pitch value is considered invalid and will not be included in the backward pitch tracking scheme of the next frame. Instead, the last valid pitch value will be used for backward pitch tracking. The pitch estimates resulting from the new tracking scheme are shown in the lower portion of Figure 3.

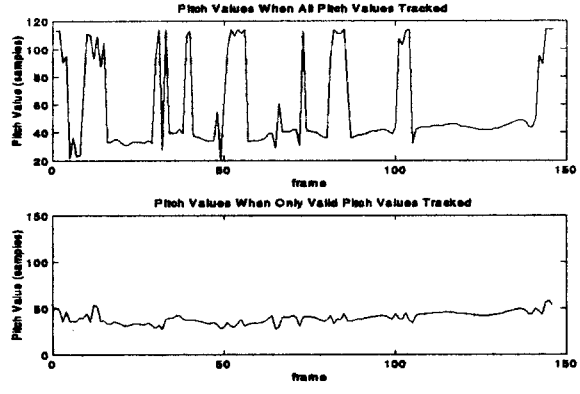

Figure 3: Comparing results of Different Pitch Tracking Schemes

Once the forward and backward pitch tracking algorithms have been employed, the initial pitch value is chosen as either the forward or backward pitch estimate. This initial pitch estimate is then refined and

a final pitch value $\hat{P}_0$ is chosen as the pitch value of the current speech frame. By finding an initial pitch value first and then refining it, computation is reduced without sacrificing accuracy. The entire pitch selection process is described in detail in [3].

3.2. Making the Voiced/Unvoiced Decisions

In a previous version of the MMBE model [3], the voicing decisions were based only on how closely the synthetic spectrum generated using the estimated pitch and spectral parameters match the original spectrum. If the spectra are closely matched in a voicing segment, the segment is considered voiced and otherwise considered unvoiced. This method of making voiced/unvoiced decisions is satisfactory when the pitch is accurately estimated. However, when errors in the pitch estimate occur, incorrect voicing decisions are made and the wrong type of excitation sequence is used to form the synthetic speech.

A prime example of the consequence of an incorrect voicing decision is shown at the beginning of the synthetic speech in Figure 4. This figure shows the

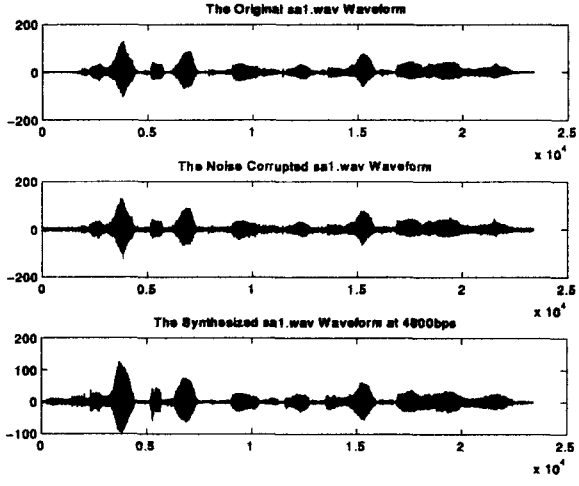

Figure 4: Consequences of Incorrect Voicing Decisions

synthetic speech resulting from a noise corrupted input signal where some speech frames were incorrectly classed as voiced. The net result is a reverberance in the synthetic speech due to its generation with a periodic excitation where it should have been generated with an unperiodic excitation.

In this MMBE model, the voicing decisions are not based exclusively on spectral mismatch. Instead, when the amount of spectral mismatch, D_k is above a threshold th_1 another set of criteria is employed such that

$$\vec{v}(k) = \begin{cases} 1 & \text{if} \begin{cases} D_k \leq th_1 \\ \text{-or-} \\ D_k \leq th_2 \quad \text{and} \quad z_c < z_{cth} \end{cases} \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

where z_c is the normalized zero crossing rate of the current voicing band and z_{cth} is the zero crossing rate threshold below which the band is considered voiced. The second spectral matching threshold, th_2 is necessary to prevent areas of silence at the beginning and end of the speech segment from being unnecessarily classified as voiced.

The zero crossing approach reduces the dependency on inaccurate pitch estimations and produces more reliable voiced/unvoiced decisions. Figure 5 shows the zero crossing rates of a speech segment and the voicing decisions both with and without the zero crossing criteria. The figure shows that although the original segment is voiced, it is considered mostly unvoiced when the zero crossing criteria is not considered. This is due to an error in the pitch estimate which inhibits the spectral matching criterion from being met. However, when the zero crossing criteria is considered, most of the segment is considered voiced and the proper type of excitation is used to generate the synthetic speech. The net result, as shown in Figure 6, is that although the synthetic speech does not perfectly match the original, it is much closer to it and the improvement is clearly audible.

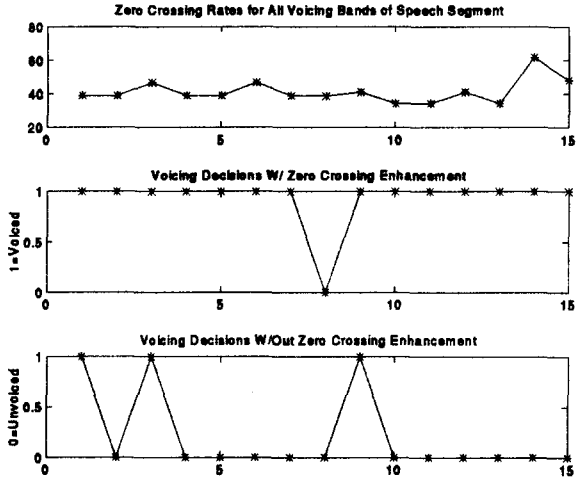

Figure 5: Comparing Results With and Without Zero Crossing Enhancement for Diphthong \c\ (all)

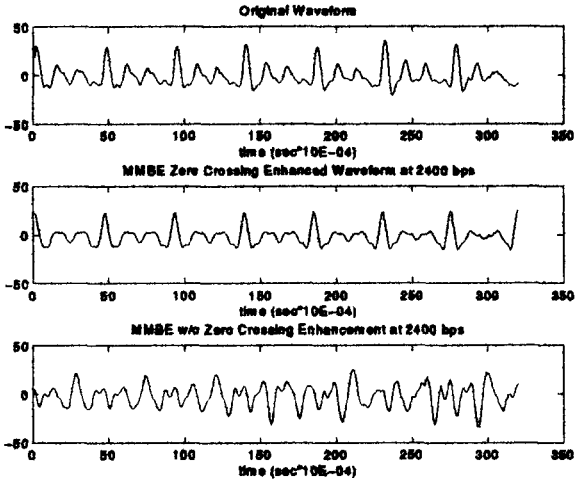

Figure 6: Waveforms of Diphthong \c\ (all)

4. PARAMETER ENCODING

Once the speech model parameter estimates have been made, the results are encoded. Table 1 shows the bit allocations for both 4800 bps and 2400 bps MMBE systems. Details on the spectral quantization and other coding techniques can be found in [3].

Table 1: Per Frame Bit Assignments of MMBE for 4800 bps and 2400 bps

Parameter Bit Rate Scheme		
Parameter	Bits:4800 bps	Bits:2400 bps
Pitch.	8	8
V/UV Decisions	K	K
Spectral Amp.	$88-K$	$40-K$
Total Bits Available	96	48

5. SPEECH SYNTHESIS

Once the speech parameters have been reconstructed, the synthetic speech signal is generated. The MMBE applies both frequency and time domain techniques to form the synthetic speech. Speech frames can have both voiced and unvoiced components. The unvoiced portion of the speech is generated in the frequency domain and then transformed into the time domain. Voiced speech segments are generated in the time domain as the sum of sinusoidal oscillators. All the voiced and unvoiced speech segments are then added together to form the synthetic speech for an entire speech frame as shown in Equation 5.

$$s(n) = s_u(n) + s_v(n) \tag{5}$$

6. CONCLUSION

An example of some synthetic speech generated using the MMBE model for a female speaker at 2400 bps is shown in Figure 7. Although testing is not yet complete, preliminary tests indicate speech quality is comparable with the MELP coder chosen as the new U.S. Federal Standard at 2400 bps for both clean and noisy speech. These results should be further improved with the refinement of the voicing decision threshold parameters and further training of the codebook used to represent the spectral shaping.

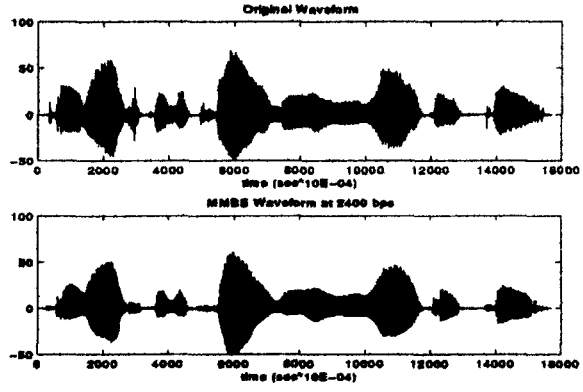

Figure 7: Comparing Original Speech to 2400 bps MMBE Synthetic Speech

1. D. Griffin, "Multiband Excitation Vocoder," *PhD thesis, Massachusetts Institute of Technology.*, Mar. 1987.

2. J. H. Hardwick, and J. Lim, "The Application of the IMBE Speech Coder to Mobile Communications", *ICASSP Proceedings.*, pp. 249-252, Jul. 1991.

3. M. L. Jamrozik, "Modified Multiband Excitation Vocoder at 2400 bps," *Master's thesis, Clemson, University.*, May. 1996.

4. J. Hardwick and J. Lim, "A 4.8 kbps multi-band excitation speech coder.", *ICASSP Proceedings.*, pp. 374–377, Sep. 1988.

5. Digital Voice Systems. "Inmarsat-M voice codec specification v3.0.", Technical report, MIT, Aug. 1991.