BAND-WIDENED HARMONIC VOCODER AT 2 TO 4 KBPS

GAO YANG*, G. ZANELLATO* & H. LEICH**

*Lernout & Hauspie Speech Products n.v.
**Lab. T.C.T.S., Faculté Polytechnique de Mons
31, Bd Dolez, 7000 Mons, Belgium

ABSTRACT

For speech coding at a bit rate below 4 kbps, the attention has been concentrated on sinusoidal-based vocoders during the past decade. Several models such as the MBE [8] have been proposed to synthesize high quality speech while removing the "buzzy" quality often produced because of over strong periodicity. This paper proposes a new model for voiced speech coding at very low bit rates, referred to as Band-Widened Harmonic coding (BWH). This model was demonstrated to be able to win some advantages over existing ones in both quality and complexity. A comparison between the BWH and the MBE will be given in this paper.

INTRODUCTION

It seems universally acknowledged that sinusoidal-based vocoder has more potential to produce higher quality speech than CELP vocoder at a bit rate below 4 kbps. In the past decade many sinusoidal vocoders have been proposed, among which MBE [8], STC [4][5], PWI [6] and TFI [7] are typical ones. This paper presents a harmonic vocoder, called Band-Widened Harmonic (BWH) vocoder, to produce high-quality speech for very low bit rates. An important point for the BWH to be different from conventional models is that controllable low-pass filtered random signals are combined into the harmonic amplitudes which are linearly interpolated between two frames; in this way the band-widths of the harmonics could be properly widened and the "buzzy" quality is thus removed.

It was found that the harmonic model is able to reconstruct voiced speech essentially perceptually indistinguishable from the original signal, if the time-variant harmonic amplitudes, frequencies and phases are correctly represented [5][6]. The harmonic amplitudes are usually linearly interpolated between two frames. The harmonic frequencies are determined by multiplying the fundamental frequency which is often evaluated by interpolating the pitch values at both ends of the frame. Accurate representation of the harmonic phases is not easy. In [5][8] one tried to represent the phase functions with a phase function model by making the phase values equal to the measured ones at the frame boundaries. However, such a method requires a relatively high bit rate. To solve this problem, the initial phases are not coded in the IMBE vocoder [9], or they are implicitly represented by the cosine and sine magnitude coefficients as in the PWI [6] and TFI [7] model. The PWI or TFI model was reported to be capable of producing high quality speech for a bit rate between 3 and 4 kbps. In these models, the frame interval for sampling the spectrum sequence is required to be short enough and a relatively large codebook is needed to correctly code the spectral magnitude coefficients because they contain the information of the phases and even the frequencies. Double number of the harmonics (cosine and sine terms), large codebook to code the magnitudes and high rate (small frame interval) for sampling the spectrum sequence all result in a high computational load. In the case that the frame interval is long for very low bit rate, a separate interpolation of the spectral magnitude and phase may yield higher performance. For a bit rate as low as 2.4 kbps, the MBE (or IMBE) model was shown to be still capable of producing high quality speech.

The aperiodic energy or wide-band harmonics may appear due to the non-linear variation of the harmonic amplitudes and frequencies in the original spectrum. The linear interpolation of the amplitudes and frequencies happens to replace the aperiodic energy with harmonics of the fundamental frequency. This causes a "buzzy" quality. The more accurate representation of the original spectrum needs a high bit rate especially for some spectral energy between periodic and aperiodic. To reduce or eliminate the "buzzy" quality, in the MBE model some bands are declared unvoiced and noisy energy produced by combining a great number of sinusoids is set in these bands. In fact, this is not an exclusive method to produce aperiodic energy. Introducing controlled random signals into the harmonic amplitudes and making the harmonic amplitudes and frequencies nonlinearly variant or time-variant could also result in aperiodic energy. Based on this idea the BWH is proposed. Another advantage of the BWH model consists in the capability of widening the band-widths of the harmonics in the region where the original spectrum is between periodic and aperiodic. This paper will compare the BWH with the IMBE.

2. NON-LINEAR INTERPOLATION OF THE FUNDAMENTAL FREQUENCY

The BWH model synthesizes voiced speech as:

$$\hat{s}(n) = \sum_{k=1}^{K} \left[1+\tilde{A}_k(n)\right] \hat{A}_k(n) \cos(\theta_k(n) + \phi_k) \qquad (1)$$

where K is the number of harmonics; the determination of the basic amplitudes $\{\hat{A}_k(n)\}$ and initial phases $\{\phi_k\}$ for each frame will be discussed in the next section; the additional terms $\{\tilde{A}_k(n)\}$, which could be zero or proper random signals depending on which frequency band the harmonic is located in, will be described in Section 4.

The phase functions $\{\theta_k(n)\}$ in (1) depend on the fundamental frequency. The precision of the time-variant fundamental frequency is important to the perceptual quality of the synthesized speech. Two commonly used phase functions are obtained by interpolating the pitch values or frequencies:

$$\bar{\theta}_k(n) = \frac{2\pi k\, n}{\tilde{\alpha}(n)\, P_0 + \alpha(n)\, P_L} \qquad (2)$$

$$\bar{\bar{\theta}}_k(n) = 2\pi k\, n \left(\frac{\tilde{\alpha}(n)}{P_0} + \frac{\alpha(n)}{P_L}\right) \qquad (3)$$

where $\alpha(n)$ linearly increases from 0 to 1 over the interpolation frame ($0 \le n < L$), and $\tilde{\alpha}(n) = 1 - \alpha(n)$; P_0 and P_L are the previous and present pitch periods, respectively. Supposing a continuous time variable, the frequency at the end of the interpolation frame is obtained by taking the derivative of the phase function:

$$\bar{\theta}_k'(L) = \frac{2\pi k}{P_L} + \frac{2\pi k\,(P_0 - P_L)}{P_L^{\,2}} \qquad (4)$$

$$\bar{\bar{\theta}}_k'(L) = \frac{2\pi k}{P_L} + \frac{2\pi k\,(P_0 - P_L)}{P_L\, P_0} \qquad (5)$$

We can see that $\bar{\theta}_k'(L)$ or $\bar{\bar{\theta}}_k'(L)$ may be quite different from the measured value $2\pi k / P_L$ when k and L are large.

In this paper the fundamental frequency function $\omega(t)$ is estimated by using the non-linear interpolation:

$$\omega(t) = \frac{(t-L/2)(t-L)\,\omega_0}{(0-L/2)(0-L)} + \frac{(t-0)(t-L)\,\omega_{L/2}}{(L/2-0)(L/2-L)} + \frac{(t-0)(t-L/2)\,\omega_L}{(L-0)(L-L/2)} \qquad (6)$$

where ω_0, $\omega_{L/2}$ and ω_L are the measured fundamental frequencies at n = 0, L/2 and L, respectively:

$$\omega_0 = 2\pi / P_0, \quad \omega_{L/2} = 2\pi / P_{L/2}, \quad \omega_L = 2\pi / P_L$$

$P_{L/2}$ is the pitch period at t = L/2, quantized with the same precision as P_L by coding the difference $P_{L/2} - P_{av}$. Here P_{av} indicates the average value:

$$P_{av} = 0.5\,(P_0 + P_L) \qquad (7)$$

Finally, we have the phase functions:

$$\theta_k(n) = k \int_0^n \omega(t)\, dt \qquad (8)$$

The performance of (8) has been tested for the interval length L=20 ms. The experiments showed that the perceptual quality with (8) is clearly better than that with (2) or (3).

3. ABOUT AMPLITUDES AND INITIAL PHASES

In the harmonic model presented by this paper, the initial phases $\{\phi_k\}$ of the harmonics are efficiently determined without using any bits. This method has been described in [1]. For the transition frame from unvoiced speech to voiced speech, the initial phases of the harmonics are specified by a random variable, of which the randomness is relevant to the pitch value. The larger the pitch value is, the smaller the randomness of the phases will be. For the successive voiced frames, the initial phases are simply determined by maintaining the continuity of the phase functions of the harmonics between the frames.

The amplitude functions $\{\hat{A}_k(n)\}$ are obtained by linearly interpolating the previous and present spectral envelope $\{\hat{A}_k\}$. In the case that the frame length is longer than 20 ms, it is better to code the spectral envelope without directly using the previous spectral envelope as an initial estimate. We synthesize the spectral envelope as

$$\hat{A}_k = g_0\,(1 + g_1\, V(k))\, H(k), \quad k = 1,2,\ldots,K \qquad (9)$$

where $\{H(k)\}$ is evaluated by slightly modifying the short-term LP spectrum with a non-linear transformation [1], which is coded using LSP parameters. The non-linear transformation in [1] has been improved again by properly maintaining the proportion between the low and high

frequency energy. The innovation vector $\{V(k)\}$ from a codebook and gains, g_0, g_1, are optimized by minimizing the sum of the squared errors between the synthetic spectrum and the ideal spectrum $\{A_k\}$:

$$\underset{V(k),g_0,g_1}{\text{Min}} \sum_{k=1}^{K} \left| A_k - g_0(1+g_1 V(k))\, H(k) \right|^2 \qquad (10)$$

Finally, the energy in the valley regions of $\{\hat{A}_k\}$ are slightly reduced in the decoder again with a similar non-linear transformation as used for $\{H(k)\}$.

4. BAND-WIDENED HARMONICS

The above methods to represent the harmonic amplitudes, frequencies and phases have efficiently reduced the "buzzy" quality. However some "buzziness" can still be heared mainly because of the linear interpolation between the previous and present spectrum. In order to inhibit the "buzzy" quality, the additional terms $\{\tilde{A}_k(n)\}$ in (1) are introduced to properly widen the band-widths of the harmonics, especially in the high frequency region. The generation of $\tilde{A}_k(n)$ is described in Fig. 1 where $H_L(z)$ is a low-pass filter and B_i is a gain factor for controlling the energy of $\tilde{A}_k(n)$. The input of the low-pass filter is a zero-mean, stable, random signal. A simple choice for $H_L(z)$ is a one-pole filter. The cut-off frequency of the low-pass filter is made approximately equal to half the fundamental frequency.

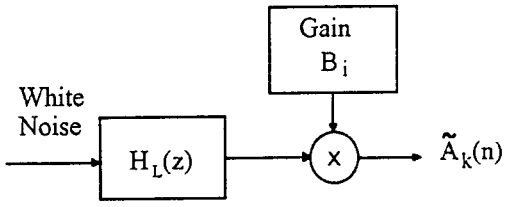


Fig. 1 The generation of the additional term $\tilde{A}_k(n)$

As in the MBE model, the spectrum could be divided into multiple frequency bands. Each band is declared strongly or faintly voiced. A strongly or faintly voiced (S/F) decision is made and a binary S/F parameter is allocated to each band. Suppose M is the number of frequency bands. The value of B_i (i=1,2,...,M) depends on the declaration of the band, whether it is strongly or faintly voiced. In general, a continous varying band dependent value of B_i would require a large number of bits to represent it accurately and make the compuation complicated. This would decrease the utility of the model in such applications as bit-rate reduction. To reduce this problem, for a strongly voiced band B_i is set to zero, otherwise B_i equals a constant $\overline{B}_i$:

$$B_i = \begin{cases} 0, & \text{strongly voiced} \\ \overline{B}_i, & \text{faintly voiced} \end{cases} \qquad (11)$$

Usually the value of $\overline{B}_i$ for a higher frequency band is larger than that for a lower frequency band, i.e. $\overline{B}_{i+1} > \overline{B}_i$. This new model differs from the MBE model in that the number M of frequency bands could be very small and even if there is no bit to code the S/F decisions the high quality could still be maintained, whereas the MBE model must use a large number of frequency bands (typically 12) and make a voiced/unvoiced (V/UV) decision to each band. In the case with few bits or without any bit to code the S/F information for the BWH model, the spectrum is still divided into several frequency bands while constant values of B_i ($B_i \neq 0$) can be determined experimentally and used for some or all of the bands. Another advantage of the BWH model consists in that the computational load of the harmonics with the additional term $\tilde{A}_k(n)$ is much less than that to generate the noisy energy by combining a great number of sinusoids.

Fig. 2 gives a comparison of the spectra of the synthetic signals with or without the additional term $\tilde{A}_k(n)$. We can see that the former is closer to the original.

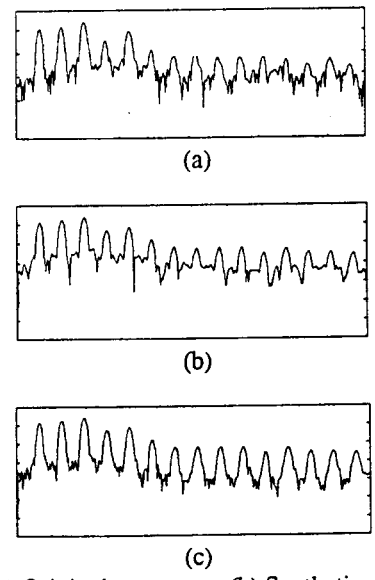


Fig. 2 (a) Original spectrum; (b) Synthetic spectrum with the additional term $\tilde{A}_k(n)$; (c) Synthetic spectrum without the additional term $\tilde{A}_k(n)$.

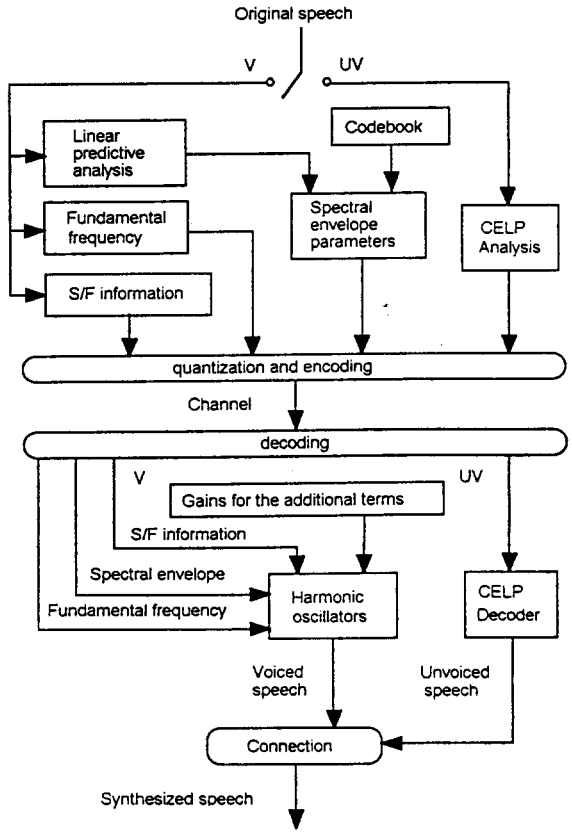
It has been reported that the number of bits to code the V/UV information in the MBE model could be reduced

assuming that if one frequency band is declared unvoiced all the higher frequency bands over that band will be always declared unvoiced. In the BWH model the same idea can be used specifying that the harmonic band-width in a higher frequency band is always larger or equal to that in a lower frequency band.

The parameters that we use in the BWH model to code voiced speech, then, are the spectral envelope, the fundamental frequency and the S/F information. Although the BWH method is aimed at voiced speech only it combines easily with other linear prediction based coders such as CELP used to code unvoiced speech, as in the PWI method. In the next section, some experimental results will be given.

5. SOME EXPERIMENTAL RESULTS

Among many applications the problem of very low bit rate for speech transmission and storage has been more and more considered. To demonstrate the performance of the Band-Widened Harmonic analysis/synthesis system, a 2.4 kbps BWH vocoder was developed. The major innovation in the BWH model is the ability to properly widen the band-widths for different frequency bands. To determine the advantage of this new model, the BWH vocoder at 2.4 to 3 kbps was compared to the MBE system at the same bit rate by changing the number of frequency bands or the number of bits to code the S/F or V/UV information.



The experiments showed that the 2.4 kbps BWH vocoder with only scalar quantization of the LP parameters (LSPs) and 0 bits to code the S/F information can produce high quality speech, the perceptual quality of which is equivalent to that by the 4.8 kbps CELP vocoder [3]. The informal tests comparing the BWH to the MBE have also been done using the same vocoder parameters for both of them except the S/F or V/UV information. It was found that the perceptual quality obtained by using the BWH vocoder is equivalent or not less than that by using the MBE vocoder, with 12 bits for both of them to code the S/F or V/UV information. When the number of the frequency bands, for which the S/F or V/UV information is coded, is reduced to 6, especially to 1 or 0, the quality of the BWH vocoder is clearly better than that of the MBE vocoder.

CONCLUSION

This paper proposes a new sinusoidal-based model for speech coding at very low bit rates. It was demonstrated that this model certainly wins some advantages over existing ones in quality and complexity.

REFERENCES

[1] Gao Yang & H. Leich, "High-Quality Harmonic Coding at Very Low Bit Rates", *ICASSP'94*, p. I-181.

[2] Gao Yang, H. Leich & R. Boite, "Voiced Speech Coding at Very Low Bit Rates Based on Forward-Backward Waveform Prediction", *IEEE Trans. on Speech and Audio Processing, January 1995.*

[3] Gao Yang, G. Zanellato and H. Leich, "A fast CELP vocoder with efficient computation of the pitch", on *EUSIPCO'92*, S6.L-7.

[4] Robert J. McAulay, & al, "Sine-Wave Amplitude Coding at Low Data Rates", *Advances in Speech Coding*, edited by B.S. Atal & al, Kluwer Academic Publisher, 1991, pp.203.

[5] Robert J. McAulay, & al, "Low-Rate Speech Coding Based on the Sinusoidal Model", *Advances in Speech Signal Processing*, edited by Sadaoki. Furui and M. Mohan Sondhi, 1991, p. 165.

[6] Kleijn, "Continous Representations in Linear Predictive Coding", *ICASSP'91*, pp. 201-204.

[7] Yair Shoham, "High-Quality Speech Coding at 2.4 to 4.0 KBPS Based on Time-Frequency Interpolation", *ICASSP'93*, p. II-167

[8] DANIEL W. GRIFFIN and JAE S. LIM, "Multiband Excitation Vocoder" *IEEE Tran. on ASSP*, Vol. 36, NO. 8, August 1988.

[9] J.C. Hardwick and JAE S. LIM, "A 4800 bps Improved Multiband Excitation Speech Vocoder" *IEEE Speech Coding Workshop*, Vancouver, B.C., Canada, 1989.