

PHASE MODELING AND QUANTIZATION FOR LOW-RATE HARMONIC+NOISE CODING

Eric W. M. Yu and Cheung-Fat Chan

Department of Computer Engineering and Information Technology
City University of Hong Kong, Tat Chee Avenue, Kowloon, Hong Kong
e-mail: e.yu@plink.cityu.edu.hk, itcfchan@cityu.edu.hk

ABSTRACT

This paper presents a novel technique for modeling and quantization of the phase information in low-rate harmonic+noise coding. In the proposed phase model, each frequency track is adjusted by a frequency deviation (FD) that reduces the error between measured and predicted phases. By exploiting the intra-frame relationship of the FD's, the phase information is represented more efficiently when compared with the representation by measured phases or by phase prediction residuals. An efficient FD quantization scheme based on closed-loop analysis is also developed. In this scheme, the FD of the first harmonic and a vector of the FD differences are quantized by minimizing a perceptually weighted distortion measure between the measured phases and the quantized phases. The proposed technique reproduces the temporal events of the original speech signal and improves the subjective quality of the synthesized speech using 13 bits per frame only.

1. INTRODUCTION

The harmonic+noise coders [1]-[6] allow mixing of voiced and unvoiced components of a speech segment in a more flexible manner when compared with the conventional coders. Hence, the harmonic+noise coders outperform the conventional CELP type coders at bit rates below 4.8 kbps. The harmonic+noise model is characterized by the fundamental frequency (pitch), harmonic magnitudes, harmonic phases, and voiced/unvoiced mixing information of a speech segment. Since coarse coding of the harmonic phases can result in poor speech quality, a lot of information bits of a low-rate harmonic+noise coder are required for phase coding. On the other hand, without encoding of the harmonic phases of the original speech signal, it was demonstrated that acceptable quality of speech can still be obtained by using the harmonic+noise model provided the frequency tracks that join the harmonics of speech segments are smooth and continue [2][4]. Therefore, due to the limited number of available bits, the phase information is not coded in general in low-rate harmonic+noise coding. In these coders, the harmonic phases are regenerated in the decoder for speech synthesis. A common approach to regenerate the phase information is to predict the harmonic phases based on the harmonic frequencies of the previous and present speech segments. Nevertheless, it was shown that the human auditory system is able to discriminate changes in the phase spectra [7][8]. Perceptual differences between the original and synthesized speech can be resulted especially at low frequencies due to the incorrect relative phase between harmonics. Therefore, the use of an efficient phase coding technique in harmonic+noise coding should improve the naturalness and quality of speech at low bit rates. Various approaches are employed for coding of the phase information in recent harmonic+noise coders. However, these phase coding schemes require in general more than 20 bits per speech segment [1][3][9]. In this paper, a new technique that leads to significantly less number of bits for phase coding is presented. The proposed technique improves the speech quality of low-rate

harmonic+noise coders by reproducing the original phase information.

2. PREDICTIVE PHASE MODEL

In harmonic+noise coding, the voiced speech $s_v(n)$ is synthesized as the sum of a set of harmonically related sinusoids

$$s_v(n) = \sum_{m=1}^M A_m(n) \cos[\theta_m(n)] \quad n = 0, 1, \dots, N-1 \quad (1)$$

where M is the total number of harmonic components in the present segment and N is the synthesis frame size. The time-varying magnitude function $A_m(n)$ in (1) is obtained by linear interpolation between the harmonic magnitudes $A_m(0)$ and $A_m(N)$. The time-varying phase function of the m^{th} harmonic is defined as

$$\theta_m(n) = \begin{cases} \phi_m, & n = 0 \\ \phi_m + \sum_{i=1}^n \omega_m(i), & n = 1, 2, \dots, N \end{cases} \quad (2)$$

where ϕ_m is the initial phase at time $n = 0$. In conventional harmonic+noise coders that do not encode the phase information, the time-varying frequency track $\omega_m(i)$ in (2) is approximated by the linear interpolation

$$\tilde{\omega}_m(i) = m\tilde{\omega}_0 + \frac{m(\omega_0 - \tilde{\omega}_0)}{N} i \quad i = 0, 1, \dots, n \quad (3)$$

where ω_0 and $\tilde{\omega}_0$ are the fundamental frequency of the present segment and the fundamental frequency of the previous segment, respectively. Combining (2) and (3) with the assumption $\omega_m(i) = \tilde{\omega}_m(i)$, we get a quadratic function of the predicted phases

$$\tilde{\theta}_m(n) = \frac{mn(n+1)}{2N} (\omega_0 - \tilde{\omega}_0) + nm\tilde{\omega}_0 + \phi_m \quad n = 0, 1, \dots, N-1 \quad (4)$$

This predicted phase function is used for smooth interpolation of the frequency track in the time domain. The phase prediction error is not minimized. Therefore, perceptual differences between original and synthesized speech are introduced.

3. PROPOSED PHASE MODEL

In order to improve the speech quality by reducing the phase prediction error, a frequency deviation (FD) is introduced to each frequency track. Let us denote the FD of the m^{th} frequency track by δ_m . Then, the proposed linear frequency track is defined as

$$\omega_m(i) = m\tilde{\omega}_0 + \frac{m(\omega_0 - \tilde{\omega}_0) + \delta_m}{N} i \quad i = 0, 1, \dots, n \quad (5)$$

Combining (2) and (5), we get an expression of the phase track

$$\theta_m(n) = \frac{\delta_m n(n+1)}{2N} + \tilde{\theta}_m(n) \quad n = 0, 1, \dots, N-1 \quad (6)$$

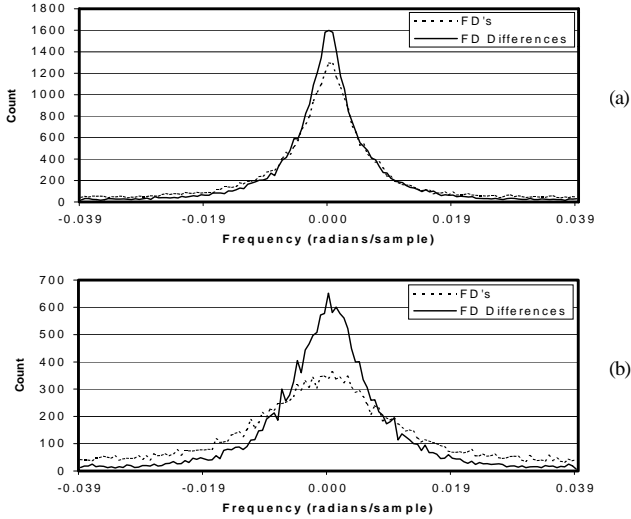


Figure 1 Histograms of FD's and FD differences in the frequency ranges of (a) 0 – 1 kHz and (b) 1 – 2 kHz of the voiced speech.

It is observed from (6) that a close approximation of the original phase track can be obtained by adjusting the predicted phase track $\tilde{\theta}_m(n)$ using the parameter δ_m . The value of δ_m is chosen such that the measured phase of the m^{th} harmonic at time $n=N$ is equal to the phase $\theta_m(N)$ evaluated using (6). Hence, δ_m can be expressed as

$$\delta_m = \frac{2}{N+1} \left[2\pi K_m + \psi_m - \tilde{\theta}_m(N) \right] \quad (7)$$

where ψ_m is the measured phase of the m^{th} harmonic at time $n=N$ and K_m is an integer for phase unwrapping.

Given the initial phase at time $n=0$ and the fundamental frequencies of the present and previous frames, the predicted phases $\{\tilde{\theta}_m(N)\}$ can be evaluated using (4). Hence, it is apparent from (7) that the FD's are equivalent representation of the measured harmonic phases in harmonic+noise synthesis. When compared with the measured harmonic phases, the FD's have a relatively smaller dynamic range and hence they are more desirable for quantization. Experimental results show that, in the frequency range of 0 to 2 kHz of the voiced speech, the variance of FD's is about 35% of that of the measured harmonic phases while the speech signal is sampled at 8 kHz and the frame size $N=160$ samples. Therefore, for a given fixed signal-to-quantization noise ratio, the number of bits required for the quantization of the FD's is less than that required for the corresponding measured harmonic phases.

Further reduction of the number of bits for phase coding can be achieved by exploiting the intra-frame relationship of the FD's. Let us define the difference between the FD's of 2 adjacent harmonics as

$$e_m = \delta_{m+1} - \delta_m \quad (8)$$

By combining (4), (7), and (8), we obtain an alternative expression of the FD difference

$$e_m = \frac{2}{N+1} \left[2\pi(K_{m+1} - K_m) + (\psi_{m+1} - \psi_m) - (\phi_{m+1} - \phi_m) - N \frac{\omega_0 + \tilde{\omega}_0}{2} - \frac{\omega_0 - \tilde{\omega}_0}{2} \right] \quad (9)$$

It is observed from (9) that the FD difference exploits both intra- and inter-frame relationships of the measured harmonic phases. By

comparing the FD's and FD differences, as given by (8), in the frequency range of 0 to 2 kHz of the voiced speech, we observe that the variance of the FD differences is about 57% of that of the FD's only. Hence, the FD differences have smaller dynamic range and the quantization of them requires less number of bits. Histograms of the FD's and FD differences in 2 different frequency ranges of the voiced speech are shown in Figure 1. As illustrated in Figure 1, both the FD's and FD differences are more concentrated around zero for low frequencies. However, the FD differences are relatively more concentrated in both frequency ranges and hence they are the more desirable parameters for quantization. It is also observed from Figure 1 that the superior statistical characteristic of FD differences is more distinct in the frequency range of 1 to 2 kHz where the variance of the FD differences is about 50% of that of the FD's. Therefore, the FD differences are more efficient representation of the phase information than the measured phases or the phase prediction residuals. Given the FD of the first harmonic, i.e. δ_1 , and the set of FD differences $\{e_m\}$, the whole set of FD's of the speech segment can be reconstructed by

$$\delta_m = \begin{cases} \delta_1, & m = 1 \\ \delta_1 + \sum_{i=1}^{m-1} e_i, & m > 1 \end{cases} \quad (10)$$

The quadratic phase tracks that are close approximations of the originals can then be obtained using (6). Note that the representation of measured harmonic phases using δ_1 and $\{e_m\}$ is lossless. Besides the distortion that can be arisen from the assumption of quadratic phase tracks in the proposed phase model, another major source of distortion in the proposed technique is the error in parameter quantization. In the following, a scheme for efficient quantization of δ_1 and $\{e_m\}$ of each speech segment is described.

4. EFFICIENT QUANTIZATION OF THE FREQUENCY DEVIATIONS

Since the human auditory system is more sensitive to the harmonics at lower frequency region than that at higher frequency region, the effects of phase distortion are more noticeable at lower frequency region. In addition, the high-frequency components of a speech signal are generally noise-like and the energy of the high-frequency harmonics is usually low. Thus, if the frequency tracks of the high-frequency harmonics are evolved smoothly, the effect on quality of the synthesized speech due to the lack of original phase information of the high-frequency harmonics is negligible. Therefore, only the phase information of the low-frequency harmonics is required to be quantized. Due to the limited number of bits in low-rate speech coding, we decided to quantize the phase information of the first 10 harmonics only.

4.1. Quantization Based on Closed-Loop Analysis

In the proposed quantization scheme, the FD of the first harmonic and the 9-dimensional vector of the FD differences are quantized. In order to minimize the quantization error of the FD's with as few bits as possible, the quantization error is minimized by a closed-loop procedure. A block diagram of the proposed FD quantization scheme is shown in Figure 2. The 10-dimensional vector of original FD's is denoted by $\delta = [\delta_1 \ \delta_2 \ \dots \ \delta_{10}]$. The table $\mathbf{t} = \{\hat{\delta}_1(j)\}_{j=0}^{J_1-1}$ stores the candidates of the FD of the first harmonic, where J_1 is the table size. The codebook $\mathbf{C} = \{\hat{\mathbf{e}}(k)\}_{k=0}^{K_1-1}$ stores the candidates of the FD difference vector

$\mathbf{e} = [e_1 \ e_2 \ \dots \ e_9]$, where K_c is the codebook size and the k^{th} candidate vector is defined as $\hat{\mathbf{e}}(k) = [\hat{e}_1(k) \ \hat{e}_2(k) \ \dots \ \hat{e}_9(k)]$. The entries of table \mathbf{t} and codebook \mathbf{C} are first concatenated to form $J_t K_c$ possible 10-dimensional vectors which are defined as

$$\mathbf{y}(\mathbf{a}_l) = [y_1(\mathbf{a}_l) \ y_2(\mathbf{a}_l) \ \dots \ y_{10}(\mathbf{a}_l)] \quad l = 0, 1, \dots, J_t K_c - 1 \quad (11)$$

$$\text{where} \quad y_i(\mathbf{a}_l) = \begin{cases} \hat{\delta}_1(j), & i = 1 \\ \hat{e}_{i-1}(k), & i = 2, 3, \dots, 10 \end{cases} \quad (12)$$

and $\mathbf{a}_l = [j \ k]$ is a vector which consists of the indices of the l^{th} possible combination of the table and codebook entries. The vector $\mathbf{y}(\mathbf{a}_l)$ is then applied to excite a recursive 1st-order system. This recursive system is described by the difference equation

$$\hat{\delta}_i(\mathbf{a}_l) = \hat{\delta}_{i-1}(\mathbf{a}_l) + y_i(\mathbf{a}_l) \quad i = 1, 2, \dots, 10 \quad (13)$$

with initial condition $\hat{\delta}_0(\mathbf{a}_l) = 0$. The output of this recursive system with respect to the input $\mathbf{y}(\mathbf{a}_l)$ is denoted by the vector

$$\hat{\delta}(\mathbf{a}_l) = [\hat{\delta}_1(\mathbf{a}_l) \ \hat{\delta}_2(\mathbf{a}_l) \ \dots \ \hat{\delta}_{10}(\mathbf{a}_l)] \quad (14)$$

This vector is a possible candidate of the FD vector $\hat{\delta}$. From the vector $\hat{\delta}(\mathbf{a}_l)$, we can re-synthesize the corresponding m^{th} harmonic phase by

$$\hat{\psi}_m(\mathbf{a}_l) = \frac{m(N+1)}{2} \left(\omega_0 - \tilde{\omega}_0 + \frac{\hat{\delta}_m(\mathbf{a}_l)}{m} \right) + Nm\tilde{\omega}_0 + \phi_m \quad (15)$$

where the initial phase ϕ_m equals the quantized phase of the previous frame. The errors due to quantizations of $\hat{\delta}_l$ and \mathbf{e} are minimized jointly by the minimization of a distortion measure between the measured harmonic phases and the re-synthesized harmonic phases. If the distortion measure is denoted as $d[\boldsymbol{\psi}, \hat{\boldsymbol{\psi}}(\mathbf{a}_l)]$, where $\boldsymbol{\psi} = [\psi_1 \ \psi_2 \ \dots \ \psi_{10}]$ is the vector of the measured phases and $\hat{\boldsymbol{\psi}}(\mathbf{a}_l) = [\hat{\psi}_1(\mathbf{a}_l) \ \hat{\psi}_2(\mathbf{a}_l) \ \dots \ \hat{\psi}_{10}(\mathbf{a}_l)]$ is the vector of the re-synthesized harmonic phases based on the index vector \mathbf{a}_l , the vector of optimum indices of the table \mathbf{t} and codebook \mathbf{C} is obtained by

$$\mathbf{a} = \arg \min_{\mathbf{a}_l} \{d[\boldsymbol{\psi}, \hat{\boldsymbol{\psi}}(\mathbf{a}_l)]\} \quad (16)$$

Then, the output of the recursive 1st-order system to the vector $\mathbf{y}(\mathbf{a})$ is the optimum quantized FD vector. In other words, if $\mathbf{a} = [j_o \ k_o]$, the optimum quantized FD vector can be reconstructed by substitution of $\hat{\delta}_1 = \hat{\delta}_1(j_o)$ and $e_i = \hat{e}_i(k_o)$ into (10). Let us denote the optimum quantized FD of the m^{th} harmonic by $\hat{\delta}_m$. A close approximation of the original phase track is

$$\hat{\theta}_m(n) = \frac{mn(n+1)}{2N} \left(\omega_0 - \tilde{\omega}_0 + \frac{\hat{\delta}_m}{m} \right) + nm\tilde{\omega}_0 + \phi_m \quad n = 0, 1, \dots, N-1 \quad (17)$$

and the optimum quantized phase of the m^{th} harmonic is $\hat{\psi}_m = \hat{\theta}_m(N)$.

4.2. Perceptually Weighted Distortion Measure

The distortion measure that we have employed in the proposed phase quantization scheme is a weighted mean-squared error function defined as

$$d[\boldsymbol{\psi}, \hat{\boldsymbol{\psi}}(\mathbf{a}_l)] = [\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}(\mathbf{a}_l)]^T \mathbf{W} [\boldsymbol{\psi} - \hat{\boldsymbol{\psi}}(\mathbf{a}_l)] \quad (18)$$

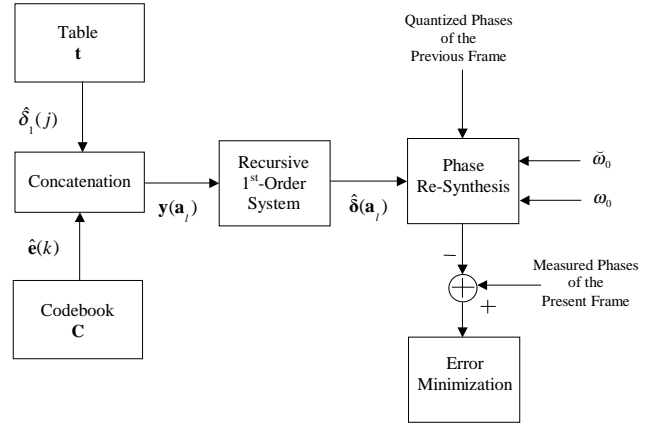


Figure 2 The proposed FD quantizer.

where \mathbf{W} is a diagonal weighting matrix used to emphasize the errors on perceptually important regions of a speech spectrum. The weighting matrix \mathbf{W} is defined as

$$\mathbf{W} = \text{diag}(v_1^2 \ v_2^2 \ \dots \ v_{10}^2) \quad (19)$$

where v_i is the adaptive weight assigned to the i^{th} harmonic phase.

As the spectral peaks of a speech spectrum represent the formants, the distortion near the spectral peaks is perceptually more noticeable than that in the lower magnitude regions. Therefore, the weights on phases of the high-magnitude harmonics should be more than the weights on phases of the low-magnitude harmonics. In addition, emphasizing the errors on phases of the high-magnitude harmonics can optimize the signal-to-quantization noise ratio of the synthesized speech signal. On the other hand, the unvoiced speech signal is noise-like. It does not exhibit the quasi-periodic nature as voiced speech signal. Therefore, the phase information of unvoiced bands is not important. In this work, the perceptual weights are determined adaptively based on both the magnitude and voiced/unvoiced decision of each harmonic band. The weight v_i is defined as

$$v_i = \begin{cases} 0.5 + \log_{10}(B_i), & B_i \geq 1; \text{ band } i \text{ is voiced} \\ 0, & B_i < 1; \text{ band } i \text{ is unvoiced} \\ 0.5, & \text{otherwise} \end{cases} \quad (20)$$

where $\{B_i\}_{i=1}^{10}$ are the magnitudes of the first 10 harmonics. As defined in (20), if the harmonic magnitude of a voiced band is high enough, the weight assigned to the corresponding phase will be an increasing function of the harmonic magnitude. For unvoiced harmonic bands which have very low magnitudes, the weights are set to zero. In order to avoid artifacts due to incorrect identification of a voiced band as unvoiced band, the weight is forced to 0.5 if the harmonic magnitude is higher than 1. On the other hand, if the harmonic magnitude of the voiced band is too low, the weight is also forced to 0.5 which is an arbitrary lower limit of the weight for voiced band.

5. PERFORMANCE EVALUATION

The proposed phase coding technique is evaluated at the rate of 13 bits/frame. The number of bits per frame assigned for quantization of the FD of the first harmonic is 5 while the number of bits per frame assigned for quantization of the FD difference vector is 8.

In Figure 3, the magnitude and phase spectra of a speech segment are shown. The quantized phase spectrum obtained using the proposed technique and the corresponding phase errors are

also shown in Figure 3. It is observed that the phase errors achieved by the proposed technique are very small.

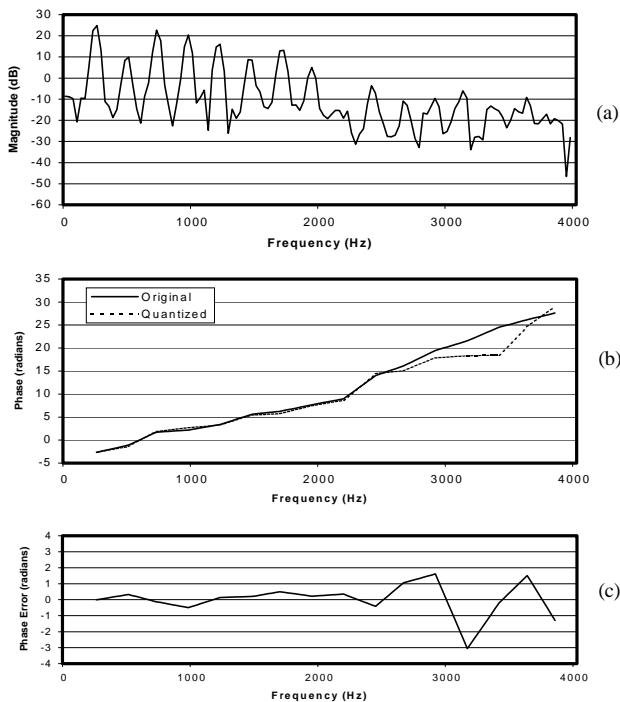


Figure 3 (a) A female speech spectrum, (b) the original and quantized phase spectra (unwrapped), and (c) the phase error.

The proposed phase coding technique has been evaluated using a speech analysis/synthesis system based on harmonic+noise model. Besides the phase information, all the parameters of the model are not quantized. Some waveforms of the original and the synthesized speech are shown in Figure 4. It is observed that the waveform shapes and temporal events of the original speech are preserved by the proposed technique.

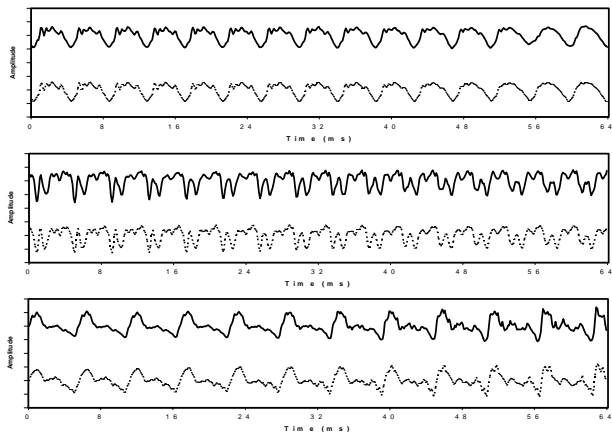


Figure 4 Waveforms of the original (solid) and the synthesized (dotted) speech obtained based on the proposed technique.

To demonstrate the viability of the proposed technique in low-rate speech coding, the proposed phase coding method at the rate of 13 bits/frame is used to replace the conventional predictive phase model of a harmonic+noise coder which was developed earlier [10]. The results of informal listening tests show that the speech synthesized using the proposed phase model is more preferable than the speech synthesized using the conventional

predictive phase model. Specifically, the naturalness and intelligibility of the coded speech are improved. A subjective A-B preference test was conducted to compare this new coder with a 2.4 kbps sinusoidal transform coder (STC) [4] that estimates the phase information in the decoder based on minimum-phase assumption. In this test, a total of 16 sentences which were contributed by 8 male and 8 female speakers were coded by the 2 different coders. The coded speech sentences were compared by 10 listeners. The results of the A-B preference test are tabulated in Table 1. From the results of this test, we observe that speech quality of the proposed coder at 2.05 kbps is comparable to the 2.4 kbps STC coder.

	Female	Male	Total
STC Coder at 2.4 kbps	36.3%	36.3%	36.3%
Proposed Coder at 2.05 kbps	40.0%	35.0%	37.5%
No Preference	23.7%	28.7%	26.2%

Table 1 Results of the A-B preference test.

6. CONCLUSION

An efficient technique for phase coding in low-rate harmonic+noise coding is proposed. In this work, the phase information is represented efficiently by taking advantage of the intra- and inter-frame relationships of the harmonic phases. A closed-loop procedure is proposed for efficient quantization of the phase information. In this quantization scheme, a perceptually weighted distortion measure between the measure and the quantized phases is minimized. The proposed technique requires only 13 bits per frame for phase representation. It is shown that the naturalness and intelligibility of the synthesized speech are improved by the use of proposed phase coding technique.

REFERENCES

- [1] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust., Speech, Signal Processing*, vol. 36, pp. 1223-1235, August 1988.
- [2] M. S. Brandstein, P. A. Monta, J. C. Hardwick, and J. S. Lim, "A real-time implementation of the improved MBE speech coder," in *Proc. IEEE ICASSP'90*, pp. 5-8.
- [3] J. S. Marques and A. J. Abrantes, "Hybrid harmonic coding of speech at low bit-rates," *Speech Commun.*, vol. 14, pp. 231-247, June 1994.
- [4] R. J. McAulay and T. F. Quatieri, "Sinusoidal coding," in *Speech Coding and Synthesis*, W. B. Kleijn and K. K. Paliwal, eds., Elsevier Science B. V., 1995, ch. 4.
- [5] A. V. McCree and T. P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech and Audio Processing*, vol. 3, pp. 242-250, July 1995.
- [6] M. Nishiguchi, A. Inoue, Y. Maeda, and J. Matsumoto, "Parametric speech coding – HVXC at 2.0 - 4.0 kbps," in *Proc. IEEE Speech Coding Workshop*, pp. 84-86, 1999.
- [7] R. D. Patterson, "A pulse ribbon model of monaural phase perception," *J. Acoust. Soc. Amer.*, vol. 82, pp. 1560-1586, 1987.
- [8] H. Pobloth and W. B. Kleijn, "On phase perception in speech," in *Proc. IEEE ICASSP'99*, pp. 29-32.
- [9] S. Ahmadi and A. S. Spanias, "A new phase model for sinusoidal transform coding of speech," *IEEE Trans. Speech Audio Processing*, vol. 6, pp. 495-501, September 1998.
- [10] E. W. M. Yu and C. F. Chan, "Harmonic+noise coding using improved V/UV mixing and efficient spectral quantization," in *Proc. IEEE ICASSP'99*, pp. 477-480.