

Harmonic vector excitation coding of speech

Masayuki Nishiguchi*

*Audio Codec Development Department, Technology Development Group, Sony Corporation,
6-7-35 Kitashinagawa, Shinagawa-ku, Tokyo, 141-0001 Japan*

(Received 22 November 2005, Accepted for publication 17 April 2006)

Abstract: A coding algorithm for speech called harmonic vector excitation coding (HVXC) has been developed that encodes speech at very low bit rates (2.0–4.0 kbit/s). It breaks speech signals down into two types of segments: voiced segments, for which a parametric representation of harmonic spectral magnitudes of LPC residual signals is used; and unvoiced segments, for which the CELP coding algorithm is used. This combination provides near toll-quality speech at 4.0 kbit/s, and communication-quality speech at 2.0 kbit/s, thus outperforming FS1016 4.8-kbit/s CELP. This paper discusses the encoder and decoder algorithms for HVXC, including fast harmonic synthesis, time scale modification, and pitch-change decoding. Due to its high coding efficiency and new functionality, HVXC has been adopted as the ISO/IEC International Standard for MPEG-4 audio.

Keywords: Harmonic coding, VXC, HVXC, CELP, Vector quantization, MPEG-4

PACS number: 43.72.Ar, 43.72.Gy [doi:10.1250/ast.27.375]

1. INTRODUCTION

Most mobile-phone and fixed-network applications for speech communication employ code-excited linear prediction (CELP) [1], which is also known as vector excitation coding (VXC) [2]. It encodes at bit rates of around 4–16 kbit/s. When the bit-rate is reduced below 4.0 kbit/s, however, the speech quality deteriorates due to the nature of waveform matching in the CELP algorithm, which requires precise phase reconstruction.

On the other hand, a class of sinusoidal coding methods, such as harmonic coding [3], multi-band excitation (MBE) [4], and sinusoidal transform coding (STC) [5], has been developed in which spectral magnitude and phase are coded separately. The harmonic structure of the power spectrum is extracted from the input speech, and the spectral magnitudes, and possibly the phase information, at harmonic frequencies are transmitted. One of the most important features of these parametric coders is that the decoder employs sinusoidal synthesis. This involves the generation of sinusoidal waveforms at harmonic frequencies by a set of oscillators, and adding them all together to make an excitation waveform. Due to this synthesis, quantization errors that arise during the coding process appear as modifications of the power spectrum, rather than as “noise.” As a result, voiced waveforms are relatively

smooth, in spite of the large quantization errors introduced because of the low bit rate of the coding. For unvoiced segments and mixed-voiced segments, the MBE coder adds band-pass noise, and STC uses phase randomization, during sinusoidal synthesis. However, when high-quality speech reproduction is required, these coders have to transmit phase information as well, which boosts the bit rate to around 6–8 kbit/s.

To further reduce the bit rate, while ensuring that speech sounds natural, a combination of the sinusoidal representation of LPC residual signals and the efficient coding of LPC parameters has been devised. The mixed-excitation LPC vocoder (MELP) [6], waveform interpolation (WI) [7], and harmonic vector excitation coding (HVXC) [8] belong to this category and provide good speech quality, comparable to or better than that of 4.8-kbit/s CELP, at coding rates of around 2.0–2.4 kbit/s. MELP and WI transmit the power spectrum of a one-pitch period of the speech waveform. MELP mixes random noise with a periodic excitation signal to generate mixed-voiced and unvoiced signals, depending on the voicing strength of sub-band signals detected by the encoder. WI does not make a decision about whether signals are voiced or unvoiced (V/UV decision). Instead, it has two types of excitation signals: one that changes rapidly, and one that changes slowly, in the time sequence of the power spectrum. These two types of signals are mixed for use by the decoder.

*e-mail: Masayuki.Nishiguchi@jp.sony.com

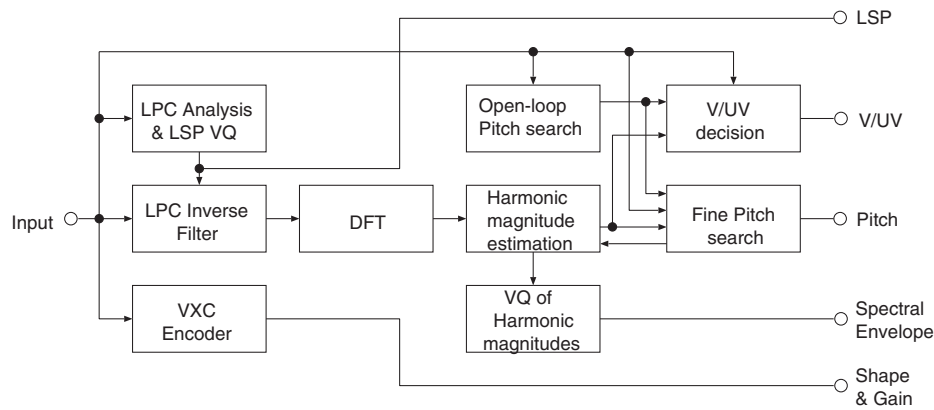


Fig. 1 Overall structure of encoder.

HVXC employs two coding algorithms: one for voiced, and one for unvoiced segments. For voiced segments, it uses the harmonic coding of the spectral magnitudes obtained by the discrete Fourier transform (DFT) of LPC residual signals. This scheme discards the phase information of LPC residuals, thus lowering the bit rate. Variable-dimension weighted vector quantization [9] is applied to spectral magnitudes, which enables them to be precisely reconstructed. For unvoiced segments, HVXC employs CELP (VXC); so it could be said to partially use waveform coding. Here, CELP uses only stochastic codebooks, and does not use a dynamic codebook at all.

It is possible for the decision-making system of a coder to make a wrong decision, and the encoder can sometimes make an error in determining whether a segment is voiced or unvoiced. The speech quality becomes unstable if the V/UV decisions are not 100% accurate. To prevent this, it is desirable to use a single coding algorithm when there are sufficient bits available, and not to have to make a decision about which algorithm to use or to switch between them. For very low-bit-rate coders, however, changing the coding algorithm depending on the nature of the input speech segment greatly improves the quality of coded speech, undoubtedly because different features of the signal should be retained in different types of segments. For voiced segments, the important points are precise representation of the power spectrum and high-frequency resolution; but for unvoiced segments, high resolution in the time domain, which allows the accurate representation of rapid changes in energy, is more important. This is why HVXC employs the combined scheme of harmonic coding and VXC, which is the major structural difference from other parametric coders like MELP and WI.

A previous paper [14] described some of the important features of the quantizer scheme in HVXC, such as variable-dimension vector quantization for harmonic spectral magnitudes, a fast codebook search, and the codebook training algorithm. This paper focuses on the overall

structure of HVXC, and some features in the decoding algorithms, which are not addressed in [14], such as a fast harmonic-synthesis algorithm and speed/pitch change algorithms. All these features distinguish HVXC from other parametric coders, providing not only high coding efficiency with reasonable complexity, but also useful functionalities like the fast playback of recorded content.

This paper is organized as follows: Section 2 presents an overview of the encoder scheme and descriptions of the individual tools. Section 3 describes the operation of the decoder and the key features mentioned above. Finally, Section 4 presents a performance evaluation, including the algorithmic delay, implementation complexity, and the results of subjective listening tests conducted as part of the MPEG-4 standardization process.

2. HVXC ENCODER

This section presents an overview of the HVXC encoder and detailed descriptions of some key components. Figure 1 shows the overall structure of the encoder [15]. Table 1 shows the bit allocation for 2- and 4-kbit/s MPEG-4 HVXC. The parameters followed by “enh” are

Table 1 Bit allocation of HVXC at 2 and 4 kbit/s.

	Voiced	Unvoiced
LSP	18 bits/20 ms	18 bits/20 ms
LSP (enh)	8 bits/20 ms	8 bits/20 ms
V/UV	2 bits/20 ms	2 bits/20 ms
pitch	7 bits/20 ms	
spectral shape	4 + 4 bits/20 ms	
spectral gain	5 bits/20 ms	
spectral shape (enh)	32 bits/20 ms	
VXC shape		6 bits/10 ms
VXC gain		4 bits/10 ms
VXC shape (enh)		5 bits/5 ms
VXC gain (enh)		3 bits/5 ms
Total 2 kbit/s	40 bits/20 ms	40 bits/20 ms
Total 4 kbit/s	80 bits/20 ms	80 bits/20 ms

for the enhancement layer and are used only in the 4-kbit/s mode. The operation of each signal-processing block of the encoder is described below.

2.1. LPC Analysis and LSP Vector Quantization

Speech input sampled at 8 kHz is divided into overlapping frames with a length of 256 samples and an interval of 160 samples. A tenth-order LPC analysis is carried out using windowed input data over one frame. The LPC parameters are converted into line-spectral-pair (LSP) parameters [10] and are vector quantized. The quantizer is composed of a base layer and an enhancement layer. The base layer uses the Partial Prediction and Multi-stage Vector Quantization (PPM-VQ) scheme [11,12], which consists of a 2-stage VQ scheme, and has two quantization modes: Normal VQ and PPM-VQ. The 1st stage of the VQ scheme is the same for both modes and is a regular 5-bit straight vector quantizer of dimension 10. In the Normal VQ mode, the 2nd stage vector quantizer simply quantizes the quantization error of the 1st stage, and no prediction scheme is employed. In the PPM-VQ mode, the 2nd stage vector quantizer quantizes the difference between the input LSP values and the dividing points of the output of the 1st stage quantizer and the quantized LSP values of the previous frame. That is, it quantizes the value T_n :

$$T_n = C_n - [P \times Q_{n-1} + (1.0 - P) \times V_{1n}], \quad (1)$$

where

C_n : input LSP of current frame

Q_{n-1} : quantized LSP of previous frame

V_{1n} : quantized LSP of current frame in the 1st stage

P : prediction coefficient (= 0.7)

In both modes, the 2nd stage has a split-VQ scheme with codebooks having sizes of 7 and 5 bits, and dimensions of 5 and 5, respectively. The LSP values are quantized using each mode, and the mode yielding the smaller quantization error is selected for each frame. A 1-bit flag indicating which mode was selected is sent to the decoder. In this VQ scheme, the use of interframe prediction is limited to just the 2nd stage of the PPM-VQ mode. Thus, the scheme provides a good compromise between coding efficiency and robustness with regard to channel errors on a transmission line. In the enhancement layer, an additional 8-bit vector quantizer of dimension 10 is used to quantize the difference between the original LSP values and the quantized LSP values of the base layer.

LPC residual signals are then computed by inverse filtering the input data using the quantized LSP parameters.

2.2. Open-Loop Pitch Search

The open-loop pitch is estimated from the peak values of the auto-correlation of the LPC residual signals. Past and current estimated pitches are used to perform pitch tracking

so as to generate a continuous pitch contour and increase the reliability of the estimated pitch. The V/UV decision of the previous frame is also used to improve the reliability of pitch tracking.

2.3. Estimation of Harmonic Magnitude and Fine-Pitch Search

The power spectrum of the LPC residual signal is then fed into the harmonic-magnitude-and-fine-pitch estimation block, which estimates the harmonic spectral envelope (SE) of the LPC residual signal and the fine pitch [4]. In the fine-pitch estimation, the pitch lag, which is the integer produced by the open-loop pitch search, is modified by \pm a few samples with a step size of 0.25. The amplitude of a basis spectrum representing one harmonic spectrum is scaled and arranged with a spacing corresponding to the modified pitch lags. The amplitude scaling for each harmonic of the fundamental and the pitch lag are adjusted simultaneously so as to minimize the difference between the synthesized power spectrum and the actual LPC residual spectrum. This operation is described below.

An N -point Hamming window ($N = 256$) is first multiplied by the LPC residual signal of N samples. An N -point DFT is performed on this windowed sequence to obtain the LPC residual spectrum, $X(j)$ ($0 \leq j < N/2$). We define $E(k)$ ($-Nr/2 \leq k < Nr/2$) to be the basis spectrum representing one harmonic spectrum. It is obtained by r -times oversampling ($r = 8$) of the DFT spectrum of the N -point Hamming window. The harmonic magnitude, $|A_m|$, is obtained from $X(j)$ and $E(k)$ by the procedure described below. First, we define the following symbols:

Pch : pitch lag of current frame

ω_0 : pitch frequency, which equals N/Pch

$m\omega_0$: center frequency of m th harmonic

We define the indices a_m and b_m , which correspond to the DFT coefficients at the lower and upper bounds of the m th harmonic, respectively, to be

$$\begin{cases} a_m = \left\lfloor \left(m - \frac{1}{2}\right)\omega_0 \right\rfloor + 1 \\ b_m = \left\lfloor \left(m + \frac{1}{2}\right)\omega_0 \right\rfloor \end{cases}, \quad (2)$$

where $\lfloor x \rfloor$ denotes the largest integer less than or equal to x . The error, ε_m , in the estimate of the m th harmonic magnitude is then given by

$$\varepsilon_m = \sum_{j=a_m}^{b_m} (|X(j)| - |A_m| |E(\text{round}(r(j - m\omega_0)))|)^2, \quad (3)$$

where $\text{round}(x)$ is a function that returns the integer nearest x . Solving

$$\frac{\partial \varepsilon_m}{\partial |A_m|} = 0 \quad (4)$$

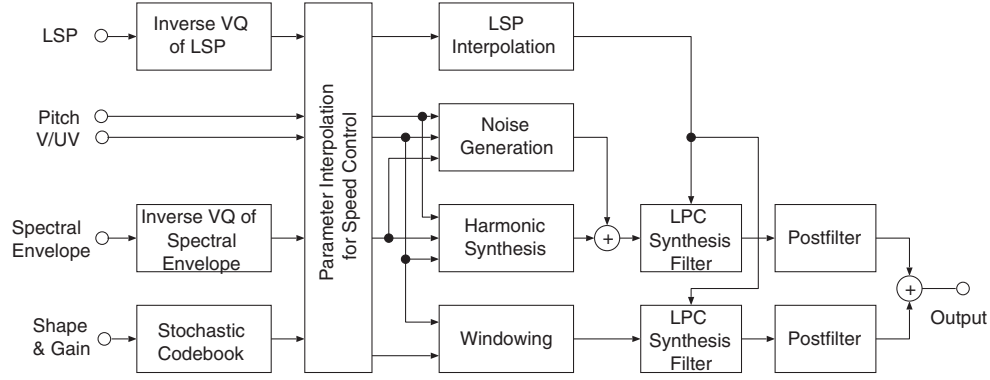


Fig. 2 Overall structure of decoder.

yields

$$|A_m| = \frac{\sum_{j=a_m}^{b_m} |X(j)| |E(\text{round}(r(j - m\omega_0)))|}{\sum_{j=a_m}^{b_m} |E(\text{round}(r(j - m\omega_0)))|^2}. \quad (5)$$

In this manner, we obtain the optimal $|A_m|$ for a given ω_0 . ε_m can then be seen as a function of only ω_0 . We call the set of harmonic magnitudes $|A_m|$ the harmonic spectral envelope (SE). Using the SE thus obtained, the total error, Er , in the estimate of all the harmonic magnitudes is

$$Er = \sum_m \varepsilon_m \quad (6)$$

for each value of the fine pitch. The fine pitch and corresponding SE that yield the minimum Er are selected for use. The SE is then vector quantized, and the fine pitch is rounded to an integer for transmission.

2.4. Vector Quantization of Harmonic Magnitudes

In order to vector quantize an SE composed of a variable number of harmonics of the fundamental, the harmonic spectral vector is first converted to a fixed-dimension (44) vector by band-limited interpolation [13]. The fixed-dimension spectral vector is then vector quantized [14]. For coding at a bit rate of 2 kbit/s, a two-stage vector quantizer, for which each stage has a 4-bit codebook, is employed for the spectral shape, together with a 5-bit scalar quantizer for the gain.

For the 4-kbit/s mode, the quantized spectral envelope from the 2-kbit/s mode, which has a fixed dimension (44), is first converted to the dimension of the original harmonics. The difference between the original harmonics and these harmonics is computed; and the resulting quantization error vector, which has the original dimension, is then quantized by additional vector quantizers. In this case, however, only the components corresponding to the lower 14 harmonics are quantized using a split VQ

scheme having codebooks with sizes of 7, 10, 9, and 6 bits and dimensions of 2, 4, 4, and 4, respectively.

2.5. Voiced/Unvoiced Decision

The normalized maximum auto-correlation of the LPC residual signals, the number of zero crossings, and the strength of the harmonic structure of the power spectrum of the LPC residual signals are used to make a decision as to whether the segment is voiced or unvoiced. These values are compared with pre-determined thresholds in parallel, and a rule-based decision is made. There are three modes for voiced segments (Mixed Voiced-1, Mixed Voiced-2, Full Voiced) and one for unvoiced segments (Unvoiced). The three voiced modes employ different amounts and bandwidths of additional noise in the harmonic synthesis, which makes mixed voiced speech sound natural and smooth (See section 3). When a segment is declared to be voiced, the decision about which of the three modes is to be used is made by comparing the value of the normalized maximum auto-correlation with pre-determined thresholds.

2.6. Vector Excitation Coding of Unvoiced Signals

For unvoiced segments, regular VXC is carried out using only stochastic codebooks. The 2-kbit/s mode employs a 6-bit shape codebook of dimension 80 and a 4-bit gain codebook. For the 4-kbit/s mode, there is an additional stage in which the quantization error of the 2-kbit/s mode is quantized again using a 5-bit shape codebook of dimension 40 and a 3-bit gain codebook.

3. HVXC DECODER

This section presents an overview of the HVXC decoder and its key features, such as fast harmonic synthesis and a pitch/speed control algorithm. Figure 2 shows the overall structure of the HVXC decoder. The basic decoding process has four steps: the dequantization of parameters; the generation of excitation signals for voiced frames by sinusoidal synthesis (harmonic synthesis) and the addition of a noise component; the generation of

excitation signals for unvoiced frames by codebook lookup; and LPC synthesis. Furthermore, a postfilter enhances the quality of synthesized speech.

For voiced frames, a fixed-dimension harmonic spectral vector obtained by dequantization of the spectral magnitude is first converted into a vector with the original dimension, which varies from frame to frame in accordance with the pitch. This is done by a dimension converter, in which a band-limited interpolator generates a set of spectral magnitudes at harmonic frequencies, without changing the shape of the SE [13]. The dimension converter in the decoder has the same structure as the one in the encoder. Based on the spectral magnitudes at harmonic frequencies, the fast harmonic-synthesis algorithm uses an inverse fast Fourier transform (IFFT) to generate a time-domain excitation signal [9] (see section 3.1). A noise component is added to make the synthesized speech sound natural. It is a spectral component of Gaussian noise, covering the frequency range of around 2–3.8 kHz, that is colored in accordance with the harmonic spectral magnitudes in the frequency domain; and its IDFT is added to voiced excitation signals in the time domain. A noise component is added when Mixed voiced-1, Mixed voiced-2, or Full voiced mode is selected; and the amount and bandwidth of the added noise is predetermined for each 2-bit V/UV value set by the encoder. Figure 3 illustrates the process. The harmonic excitation signal for voiced segments, including the added noise, is fed into the LPC synthesis filter, and then into the postfilter.

For unvoiced segments, the usual VXC decoding algorithm is used, in which an excitation signal is generated by multiplying the gain by the shape codevector. The result is

fed into the LPC synthesis filter, and then into the postfilter. Finally, the synthesized speech components for voiced and unvoiced segments are added in the time domain to form the output signal.

3.1. Fast Harmonic Synthesis by IFFT

One drawback of harmonic coding is the high complexity of the synthesizer. Assume that the voiced output, $v(n)$, is computed directly from the equation

$$v(n) = \sum_m A_m(n) \cos(\theta_m(n)) \quad (7)$$

$$(0 \leq m < M, \quad 0 \leq n < N_0),$$

where $A_m(n)$ are interpolated amplitudes and $\theta_m(n)$ are phases [4], with n being a discrete time index and m being a harmonic index. Then, the complexity of the implementation is on the order of $\gamma N_0 M$, where γ is a constant related to the interpolation of the amplitude and phase, N_0 is the frame interval (in a sample), and M is the maximum number of harmonics. Typically, $N_0 = 160$, $M = 64$, and $\gamma = 5$. In MPEG-4 HVXC [19], the complexity of this process is reduced through the use of a fast synthesis method [9] employing an IFFT and sampling-rate conversion, in which $A_m(n)$ and $\theta_m(n)$ are linearly interpolated. The procedure is illustrated in Fig. 4.

Suppose that, for the k th frame, there is a spectrum with M_1 harmonics and that the magnitude of each is $|A_m|$ ($0 \leq m < M_1$). The pitch lag expressed in terms of the number of samples is $2M_1$. Appending zeros to the array of amplitudes, $|A_m|$, yields a new array with 2^b components ranging from 0 to π . The number b can be arbitrarily chosen so that $M \leq 2^b$. The same processing is done on the array of phase data. The phase data used here are generated from those of the previous frame, based on the assumption that the fundamental frequency is linearly interpolated. This is described as phase prediction in [4]. At the onset of speech, where the previous frame is unvoiced, the array of phase data is initialized using random values uniformly distributed between 0 to 0.5π . A 2^{b+1} point IFFT is applied to the magnitude and phase arrays under the constraint that the results be real numbers.

This operation yields an over-sampled version of the time-domain waveform over a one-pitch period. Let this be $w_1(i)$ ($0 \leq i \leq 2^{b+1}$). Thus, 2^{b+1} points are used to express the one-pitch period of the waveform, whereas the actual pitch is $2M_1$ because the over-sampling ratio ov_1 is

$$ov_1 = 2^b / M_1. \quad (8)$$

Another one-pitch period of the waveform for the $(k+1)$ th frame can be similarly obtained. It has an over-sampling ratio of

$$ov_2 = 2^b / M_2, \quad (9)$$

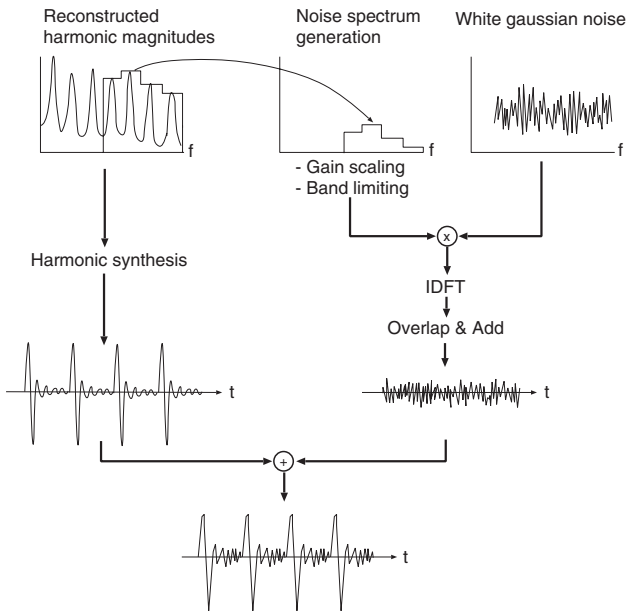


Fig. 3 Excitation generation for voiced and mixed-voiced segments.

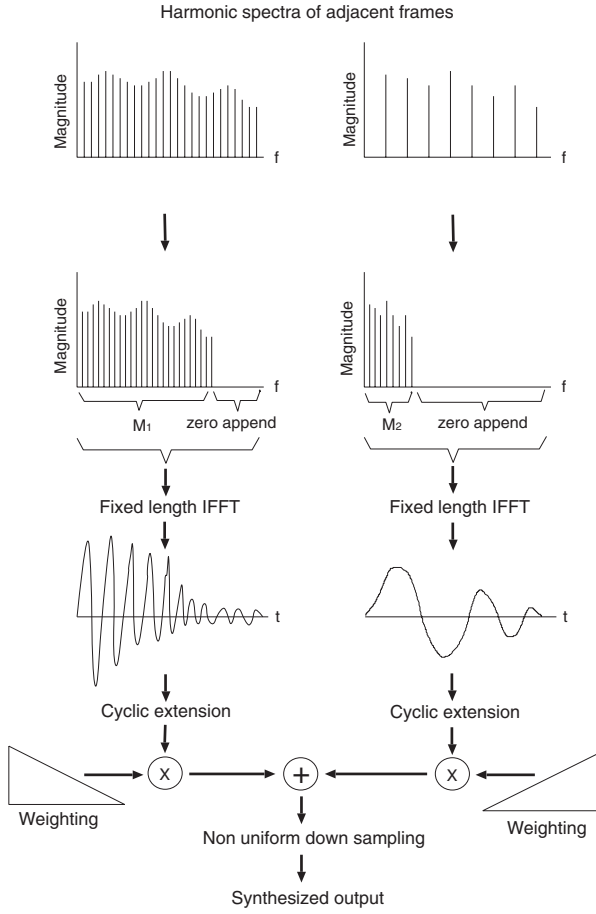


Fig. 4 Fast harmonic synthesis.

where $2M_2$ is the pitch lag. Let this waveform be $w_2(i)$ ($0 \leq i \leq 2^{b+1}$). Here, the function $f(n)$, which maps the time index, n , from the version obtained at the original sampling rate to the over-sampled version, is defined under the condition that the pitch be linearly interpolated to be

$$f(n) = \int_0^n \left(ov_1 \frac{N_0 - t}{N_0} + ov_2 \frac{t}{N_0} \right) dt. \quad (10)$$

The number of over-sampled data needed to reconstruct a waveform of length N_0 at the original sampling rate is at most L :

$$L = \text{round}(f(N_0)) = \text{round}\left(\frac{N_0}{2}(ov_1 + ov_2)\right), \quad (11)$$

By cyclically extending $w_1(i)$ and $w_2(i)$, we obtain the waveforms $\tilde{w}_1(l)$ and $\tilde{w}_2(l)$, both which have length L :

$$\tilde{w}_1(l) = w_1(\text{mod}(l, 2^{b+1})) \quad (0 \leq l < L), \quad (12)$$

$$\tilde{w}_2(l) = w_2(\text{mod}(\text{offset} + l, 2^{b+1})) \quad (0 \leq l < L), \quad (13)$$

where

$$\text{offset} = 2^{b+1} - \text{mod}(L, 2^{b+1}), \quad (14)$$

and $\text{mod}(x, y)$ returns the remainder of x divided by y .

These two waveforms, $\tilde{w}_1(l)$ and $\tilde{w}_2(l)$, from the

spectra of the k th and $(k+1)$ th frames have the same *pseudo* pitch ($= 2^{b+1}$) and are aligned. So, simply adding them together using appropriate weights produces $w(l)$:

$$w(l) = \frac{L-1}{L} \tilde{w}_1(l) + \frac{1}{L} \tilde{w}_2(l) \quad (0 \leq l < L), \quad (15)$$

where each A_m is linearly interpolated between adjacent frames. Finally, $w(l)$ has to be resampled so that the resulting waveform can be expressed on the original uniform sampling grid. This operation brings the waveform back from the *pseudo* pitch domain to the real pitch domain, as well. In principle, the resampling operation is just:

$$v'(n) = w(f(n)) \quad (0 \leq n < N_0). \quad (16)$$

Usually $f(n)$ does not return an integer value. So, $v'(n)$ is obtained by linearly interpolating $w(\lceil f(n) \rceil)$ and $w(\lfloor f(n) \rfloor)$. For a more general formulation, a higher-order interpolation could be used. $\lceil x \rceil$ and $\lfloor x \rfloor$ denote the smallest integer greater than or equal to x , and the largest integer less than or equal to x , respectively. It should be noted that Eq. (10) applies only when both the k th and $(k+1)$ th frames are voiced; otherwise, the spectral magnitude, pitch value, and over-sampling ratio of only the voiced frame are used without any interpolation.

$v'(n)$ is a good approximation of $v(n)$ in Eq. (7), and reduces the complexity of the process to the order of $\alpha 2^b(b+1) + \beta N$. α and β are constants related to IFFT and linear interpolation, respectively. Listening experiments demonstrated a b of 6 to be sufficient, and synthesized speech was indistinguishable from that obtained by the direct method of Eq. (7). The average segmental SNR between the two results was about 32 dB. Typically, $\alpha = 7$ and $\beta = 12$. So, the complexity is less than one-tenth that of the direct-synthesis method.

3.2. Time Scale Modification

One of the most important features of HVXC is its speed-control capability. The HVXC decoder has a scheme for parameter interpolation that generates the parameters at any arbitrary instant of time. A sequence of encoded parameters with modified intervals is fed into the speech synthesizer to generate speech with a modified time scale. The operation of this time-scale-modification block is described below [15,16].

The arrays of original and interpolated parameters are denoted as $prm[n]$ and $m_prm[m]$, respectively, where n and m are the time indices (frame number) before and after the time scale modification. The frame intervals are both 20 ms long. The parameters needed are pitch, LSP, and residual spectrum. We define the ratio of the speed change, s , to be

$$s = N_1/N_2, \quad (17)$$

where N_1 ($0 \leq n < N_1$) is the duration of the original speech and N_2 ($0 \leq m < N_2$) is the duration of the speed-controlled speech. The time-scale-modified parameters are

$$m_prm[m] = prm[m \times s]. \quad (18)$$

In general, however, $m \times s$ is not an integer. So, it is necessary to define

$$\begin{cases} m_L = \lceil m \times s \rceil - 1 \\ m_H = m_L + 1 \end{cases} \quad (19)$$

to generate parameters at the time index $m \times s$ by linearly interpolating the parameters at time (frame) indices m_L and m_H . In order to perform the linear interpolation, we define

$$\begin{cases} d_L = m \times s - m_L \\ d_H = m_H - m \times s \end{cases} \quad (20)$$

Then, Eq. (18) can be approximated as follows:

$$m_prm[m] = prm[m_L] \times d_H + prm[m_H] \times d_L. \quad (21)$$

This operation is shown in Fig. 5. Since the excitation signal at the VXC decoder is a time domain waveform, the interpolation method of Eq. (21) cannot be used. Consequently, we have to take one frame (160 samples) of excitation samples from the original parameters $prm[n]$ centered around the time $m \times s$, and compute the energy, E , over the frame (160 samples). Gaussian noise consisting of 160 samples is then generated, and the gain is adjusted so that it has an energy equal to E . This gain-adjusted Gaussian noise sequence is used for time-scale-modified VXC excitation, $m_prm[m]$.

Depending on the V/UV decisions at m_L and m_H , the

interpolation strategy varies in the following way:

- **voiced–voiced**

All of the parameters (LSP, pitch lag, and fixed-dimension spectral envelope) are interpolated using Eq. (21).

- **unvoiced–unvoiced**

The LSP parameters are interpolated using Eq. (21); the VXC excitation centered around $m \times s$ is generated as described above.

- **voiced–unvoiced**

If $d_L < d_H$,

then all of the parameters of frame m_L are used.

If $d_L \geq d_H$,

then the LSP parameters of frame m_H are used; the VXC excitation centered around m_H is generated in the same manner as described above.

- **unvoiced–voiced**

If $d_L < d_H$,

then the LSP parameters of frame m_L are used; the VXC excitation centered around m_L is generated in the same manner as described above.

If $d_L \geq d_H$,

then all of the parameters of frame m_H are used.

In this manner, all of the necessary parameters for the HVXC decoder are generated. Feeding these modified parameters, $m_prm[m]$, into the speech synthesizer in the same way as for the usual decoding process yields the time-scale-modified output. Clearly, when $N_2 < N_1$, decoding is speeded up; and when $N_2 > N_1$, decoding is slowed down. The power spectrum and pitch are not affected by this speed control, meaning that good-quality speech is obtained for speed-control factors in the range of about $0.5 < s < 2.0$.

3.3. Pitch Modification

Another important feature of HVXC is the pitch-modification functionality, which enables the pitch of the synthesized speech to be altered during decoding. In the regular decoding process, the dimension converter converts the fixed-dimension harmonic spectral vector into one with the original dimension, as described in [14]. Pitch modification is carried out by simply modifying the target pitch frequency for the dimension conversion [16]. Figure 6 shows an example of pitch-up decoding, in which the spacing between the harmonics (pitch frequency) is made wider than that for normal decoding, without any alteration of the shape of the SE.

4. PERFORMANCE EVALUATION AND CONCLUSIONS

This section concerns the algorithmic delay of the encoder and decoder, implementation complexity, and listening-test results for HVXC.

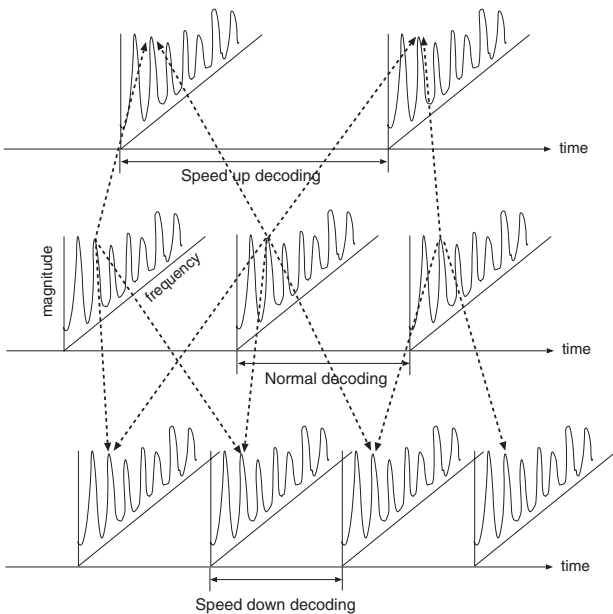


Fig. 5 Time scale modification.

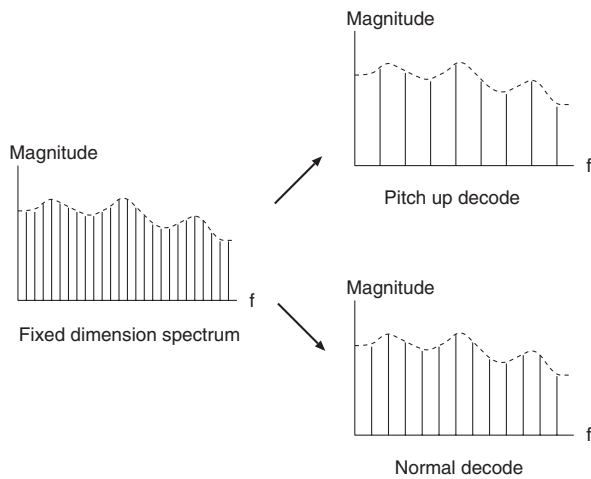


Fig. 6 Example of pitch modification by dimension conversion.

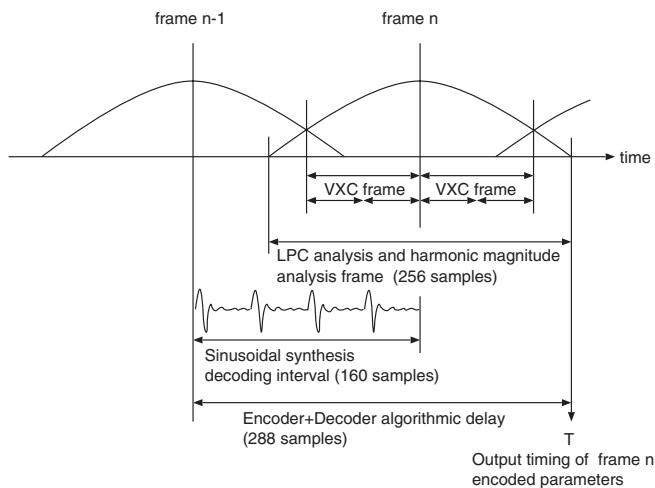


Fig. 7 Framing structure and algorithmic delay of HVXC.

4.1. Algorithmic Delay

Figure 7 shows the framing structure of the encoder and decoder process. The encoder process for the current frame can be started when the number of samples of frame data in the input buffer reaches 256. Let this time be T . It should be noted that the LPC coefficients of only the current frame are used in inverse filtering and VXC analysis to avoid the introduction of an additional look-ahead delay. The decoder process for a voiced waveform between the center of the current frame and the center of the previous frame involves sinusoidal synthesis based on the encoded parameters of the current and previous frames. This operation can be started at time T , assuming that the processing delay of the encoder is zero. The total algorithmic delay for the encoder and decoder is then equal to the period of time between the center of the previous frame and time T , which is 288 samples, or 36 ms.

Table 2 Implementation complexity of HVXC.

Table ROM	26 kbyte
Instruction ROM	30 kbyte
Data RAM	9 kbyte
MIPS	Enc: 20 MIPS Dec: 8 MIPS

4.2. Complexity

We implemented the HVXC encoder and decoder on a 16-bit fixed-point DSP with a single multiplier. The required ROM, RAM, and MIPS for full-duplex implementation are shown in Table 2. As can be seen, the memory and MIPS requirements are modest enough to enable implementation on a low-end, general-purpose DSP chip.

4.3. Listening Test Results

The performance of HVXC when tested as one candidate for the MPEG-4 standard was presented in a previous paper [14]. During the standardization process, however, some components were modified, such as the LSP quantization scheme and bit allocations for harmonic spectral magnitudes, as shown in Section 2.

In order to compare the sound quality of the finalized MPEG-4 speech coders to that obtainable with existing standards, official subjective tests were conducted in August 1998 as the final step in the MPEG-4 standardization process, using one Japanese test site and two European test sites [17]. For HVXC, 15 Japanese items were evaluated by 16 Japanese listeners at the Japanese site; and a total of 15 English, German, and Swedish items were evaluated by 18 German listeners and 16 Finnish listeners at two European sites. HVXC at 2.0 and 4.0 kbit/s, and FS1016 at 4.8 kbit/s were evaluated in the same session. ACR tests with a 5-point grading scale were used for all the tests. Most of the items used in this test were pure speech items. Some of the results obtained at one site were somewhat different from those obtained at another site. However, the overall results on coder performance obtained from the tests were consistent, irrespective of test site or language. The mean opinion scores (MOS) for a 95% confidence interval (CI) for the Japanese and European sites are shown in Fig. 8. They show that the performance of MPEG-4 HVXC at 2.0 and 4.0 kbit/s is significantly better than that of FS1016 4.8-kbit/s CELP.

The quality of time-scale-modified and pitch-modified speech was assessed at the beginning of the MPEG-4 standardization process using the coders that were initially proposed, although this was not an official ACR tests [18]. The evaluation results revealed that the quality of both types of modified speech produced with HVXC at 2.0 kbit/s were "very good"; and no impairment due to

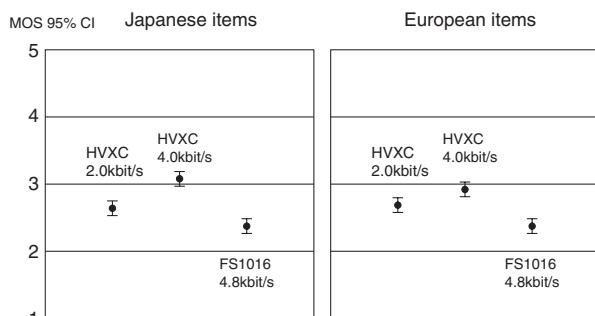


Fig. 8 Subjective test results.

the modification algorithms in Sections 3.2 and 3.3, was detected.

As described here, HVXC provides high coding efficiency and new functionality, such as time-scale modification and pitch modification, which will open the way to new multimedia applications. HVXC has already been adopted as the ISO/IEC International Standard for MPEG-4 Audio [19].

REFERENCES

- [1] M. R. Schroeder and B. S. Atal, "Code-excited linear predictive (CELP): High-quality speech at very low bit rates," *Proc. ICASSP 85*, pp. 937–940 (1985).
- [2] A. Gersho and R. M. Gray, *Vector Quantization and Signal Compression* (Kluwer Academic Publishers, 1992), pp. 621–628.
- [3] J. S. Marques, L. B. Almeida and J. M. Tribolet, "Harmonic coding at 4.8 kb/s," *Proc. ICASSP 90*, pp. I-17–20 (1990).
- [4] D. W. Griffin and J. S. Lim, "Multiband excitation vocoder," *IEEE Trans. Acoust. Speech Signal Process.*, **36**, 1223–1235 (1988).
- [5] R. J. McAulay and T. F. Quatieri, "Speech analysis/synthesis based on a sinusoidal representation," *IEEE Trans. Acoust. Speech Signal Process.*, **34**, 744–754 (1986).
- [6] A. V. McCree and T. P. Barnwell III, "A mixed excitation LPC vocoder model for low bit rate speech coding," *IEEE Trans. Speech Audio Process.*, **3**, 242–250 (1995).
- [7] W. B. Kleijn and J. Haagen, "A speech coder based on decomposition of characteristic waveforms," *Proc. ICASSP 95*, pp. I-508–511 (1995).
- [8] M. Nishiguchi, K. Iijima and J. Matsumoto, "Harmonic vector excitation coding of speech at 2.0 kbps," *IEEE Workshop Speech Coding*, pp. 39–40 (1997).
- [9] M. Nishiguchi and J. Matsumoto, "Harmonic and noise coding of LPC residuals with classified vector quantization," *Proc. ICASSP 95*, pp. I-484–487 (1995).
- [10] F. Itakura, "Line spectral representation of linear predictive coefficients of speech signals," *J. Acoust. Soc. Am.*, **57**, S35 (1975).
- [11] N. Tanaka, T. Morii, K. Yoshida and K. Honma, "A multi-mode variable rate speech coder for CDMA cellular systems," *Proc. IEEE VTC*, pp. 198–202 (1996).
- [12] M. Nishiguchi, A. Inoue, Y. Maeda, J. Matsumoto and N. Tanaka, "MPEG-4 parametric speech coding — HVXC," *Tech. Rep. IEICE*, SP98-90, Vol. 98, No. 424, pp. 27–34 (1998).
- [13] M. Nishiguchi, J. Matsumoto, S. Ono and R. Wakatsuki, "Vector quantized MBE with simplified V/UV division at 3.0 kbps," *Proc. ICASSP 93*, pp. II-151–154 (1993).
- [14] M. Nishiguchi, "Weighted vector quantization of harmonic spectral magnitudes for very low-bit-rate speech coding," *Acoust. Sci. & Tech.*, **27**, 43–49 (2006).
- [15] M. Nishiguchi, A. Inoue, Y. Maeda and J. Matsumoto, "Parametric speech coding — HVXC at 2.0–4.0 kbps," *IEEE Workshop Speech Coding*, pp. 84–86 (1999).
- [16] M. Nishiguchi, "MPEG-4 speech coding," *Proc. Audio Eng. Soc. 17th Int. Conf.*, pp. 139–146 (1999).
- [17] ISO IEC JTC1/SC29/WG11 MPEG98/N2424, "Report on the MPEG-4 speech codec verification tests," Oct. (1998).
- [18] ISO IEC JTC1/SC29/WG11 MPEG95/N1063, "Report of the Adhoc Group on the evaluation of tools for nontested functionalities of audio submissions to MPEG-4," Nov. (1995).
- [19] ISO/IEC 14496-3:1999, *Coding of Audiovisual Objects — Part 3: Audio*, Version 1, Dec. (1999).



Masayuki Nishiguchi received his B.E. degree in Physical Electronics from the Tokyo Institute of Technology in 1981 and his M.S. degree in Electrical and Computer Engineering from the University of California, Santa Barbara in 1989. Since 1981, he has been with the Sony corporation, where he is engaged in the development of speech and audio coding, and signal processing algorithms. He is currently a general

manager of the Audio Codec Development Department of Sony. He is a member of IEEE.