Synthesis and Implementation of Virtual Bass System with a Phase-Vocoder Approach*

MINGSIAN R. BAI, *AES Member,* **AND WAN-CHI LIN**
(msbai@mail.nctu.edu.tw) (ChichyLin@tp.cmcs.com.tw)

Department of Mechanical Engineering, National Chiao-Tung University, Hsin-Chu 300, Taiwan

A bass enhancement technique based on a phase-vocoder approach is presented. Instead of direct bass boosting, the proposed method creates a bass impression by exploiting the psychoacoustic properties of humans. This technique is most useful in audio reproduction using small loudspeakers that have no low-frequency capability, where direct boosting will likely result in nonlinear distortions. In light of psychoacoustics, the bass effect is synthesized by augmenting the original signals with high-frequency harmonics. Unlike conventional methods that rely on nonlinear processing, the proposed method performs the required frequency transformation by using a phase-vocoder approach. Apart from frequency transformation, another key element of the proposed technique is the magnitude adjustment of the generated harmonics. The underlying principle for magnitude adjustment is based on a polynomial model of equal-loudness contours. The method is implemented on a digital signal processor with the aid of multirate signal processing. To validate the proposed technique, objective and subjective experiments are conducted for PC multimedia loudspeakers and handset loudspeakers. The subjective listening experiment followed the procedure of multistimuli with the hidden reference and anchor (MUSHRA), and the data were analyzed by using the multianalysis of variance (MANOVA) method. As indicated by the results, the proposed technique proved effective in rendering bass impression with acceptable audio quality.

0 INTRODUCTION

Computer, communication, and consumer electronics (3C) products are facing strong demands from the marketplace for more audio and video functions. For example, game stations are designed to provide quality animations as well as audio rendering with spatial fidelity. Another example is that a third-generation (3G) handset has been devised to serve not only as a phone but as an MP3 player, a TV, and a camera. Therefore new thoughts must be given to the design of audio systems that best address the needs of 3C products.

Bass enhancement is one of the key elements in modern audio reproduction. A longstanding challenge in bass enhancement has been how to produce bass using a loudspeaker that has no low-frequency capability, such as a handset microspeaker. A common way of dealing with this problem is through the use of equalizers. The low-frequency gain is increased using shelving filers or other electronic means. This requires substantial increases in system headroom such that the system can be driven beyond its operating limits. Unless the loudspeaker and the power amplifier are redesigned completely for the low-frequency purpose, direct bass boosting will only result in nonlinear distortions, or even permanent damage.

Instead of the physical means mentioned, this paper focuses on an alternative approach that avoids these drawbacks of conventional methods. It is termed the virtual bass (VB) system because it creates bass impression psychoacoustically without actually producing low-frequency signals. The VB system exploits a bandwidth extension property of human hearing, namely, that humans are capable of "extrapolating" the missing fundamental in the low-frequency range based on higher harmonics. Even if in a harmonic complex the fundamental frequency is missing, it will still be perceived as a residue pitch, which in this case is sometimes called virtual pitch. The fact that the fundamental need not be physically present to evoke a pitch percept at the same fundamental is an attractive feature to enhance bass sounds produced by small loudspeakers. A very comprehensive coverage of psychoacoustic bandwidth extension can be found in the monograph by Larsen and Aarts [1]. MaxxBass represents one of the commercial systems that made use of this psychoacoustic property [2]. Superharmonics are created using nonlinear operations. The amplitudes of the harmonics are determined according to an approximate rule of loudness.

*Manuscript received 2006 July 6; revised 2006 September 26.

Shashoua and Glotter proposed another VB system using similar principles [3]. A residue harmonics generator in a feedback loop was used in their system to create the required harmonics. The processing of this approach was computationally expensive. Gan et al. presented a VB system based on a similar psychoacoustic principle. However, in their approach superharmonics were created by signal modulation [4].

Although the approach proposed in this paper relies on psychoacoustic principles similar to previous research, the method of realization is quite different. The harmonic generation in this method is based on the phase-vocoder approach [5]. The phase vocoder is a simple but effective technique for time–frequency processing. The phase-vocoder approach offers a significant advantage over previous approaches for harmonic generation in that phase coherence can be preserved during processing. Phase coherence plays an important role in audio quality [5]. An artifact, the so-called phasiness, will be audible if phase propagation from frame to frame is not properly dealt with in harmonic generation.

The determination of the magnitude of each harmonic is another important issue in the VB system. This is generally achieved by taking into account the loudness of the signals in low frequencies. According to psychoacoustics, human perception of the loudness of sound is strongly frequency dependent and can be characterized by equal-loudness contours [6]. For natural bass reproduction the magnitude of the superharmonics generated by the VB system should be adjusted according to the equal-loudness contour. Instead of using approximate loudness analyzers [3], [4], we use an accurate polynomial model to calculate the magnitudes of the generated harmonics with appropriate loudness.

In the implementation phase the computational cost can be reduced substantially by taking advantage of multirate processing. Because the VB system basically aims at a very low frequency range, computation can be carried out by down sampling the audio signal to a much lower rate. The processing efficiency of up or down sampling can be further improved by polyphase implementation [7], [8].

Objective and subjective experiments were conducted in order to validate the proposed VB system. The subjective test is arranged according to ITU-R BS.1116 [9]. Two-channel stereo loudspeakers for multimedia and a microspeaker of a handset were employed as the rendering transducers to compare the proposed method with the MaxxBass system. The listening test was conducted following the multistimuli with the hidden reference and anchor (MUSHRA) procedure [10]. To justify the statistical significance of the results, the data of the subjective tests were analyzed by a multianalysis of variance (MANOVA) [11].

1 FUNDAMENTALS OF THE PHASE VOCODER

1.1 Time–Frequency Processing Using Phase Vocoders

Implementation of a phase vocoder generally involves three aspects: analysis, transformation, and synthesis, as shown in Fig. 1. Analysis is to analyze the frequency content. Transformation carries out the intended time–frequency processing. Synthesis assembles the transformed frequency-domain components into a final waveform. The phase vocoder per se constitutes frame-based processing, which can be carried out efficiently with the aid of the fast Fourier transform (FFT). Using the short-time Fourier transform (STFT), a discrete-time sound signal can be represented by a two-dimensional time–frequency diagram (Fig. 2). The time–frequency distribution of the signal is modified in some ways and then a new sound is created. The STFT of a signal $x(n)$ is given by [12]

$$X(n, k) = \sum_{m=-\infty}^{\infty} x(m)h(n-m)e^{-j2\pi mk/N} = |X(n, k)|e^{j\phi(n, k)}, \quad k = 0, 1, \ldots, N-1 \tag{1}$$

where $X(n, k)$ is a complex sequence of time index n and frequency bin k and $|X(x, k)|$, $\phi(n, k)$, and $h(n-m)$ denote, respectively, the magnitude and phase of $X(n, k)$ and the impulse response of a finite-length window.

1.2 Implementation of the Phase Vocoder

There are two approaches to realize a phase vocoder: the FFT/IFFT structure and the FFT/sum-of-sinusoids structure [13]. Fig. 3 shows a FFT/IFFT structure for a phase vocoder. The STFT of the analysis stage is shifted every R_a samples in time, or the analysis hop size R_a. The STFT at the time instant $n = sR_a$ of the sth frame can be written as

$$X(sR_a, k) = |X(sR_a, k)|e^{j\phi(sR_a, k)}. \tag{2}$$

The magnitude $|X(sR_a, k)|$ and the phase $\phi(sR_a, k)$ can be modified in the frequency domain to obtain a revised spectrum $Y(sR_s, k)$, where R_s is the synthesis hop size. The modified spectrum in the frequency domain is then in-

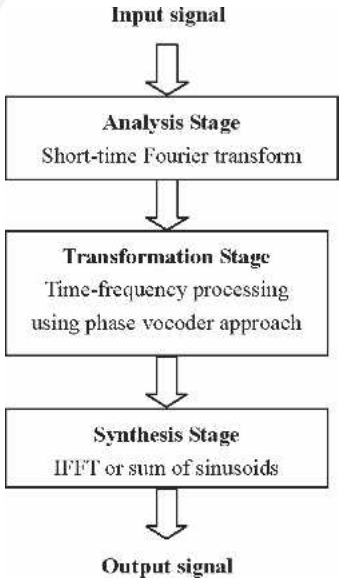


Fig. 1. Schematic diagram showing procedure of phase vocoder.

verted to the time domain by using IFFT to yield $y_s(n)$. Note that a synthesis window $f(n)$ may be needed to minimize the time-aliasing problem during IFFT. Hence the short-time segments are windowed and are overlap–added to produce a new signal,

$$y(n) = \sum_{s=-\infty}^{\infty} f(n - sR_s) y_s(n - sR_s). \quad (3)$$

Another approach of implementing a phase vocoder is the FFT/sum-of-sinusoids structure, as shown in Fig. 4. A detailed discussion can be found in [13].

Phase is a key element to the sound quality in phase-vocoder–based synthesis [14]–[16]. Phasiness may arise when phase coherence is not well maintained, and artifacts may be audible due to phase and amplitude modulation. In order to maintain phase coherence, a phase unwrapping

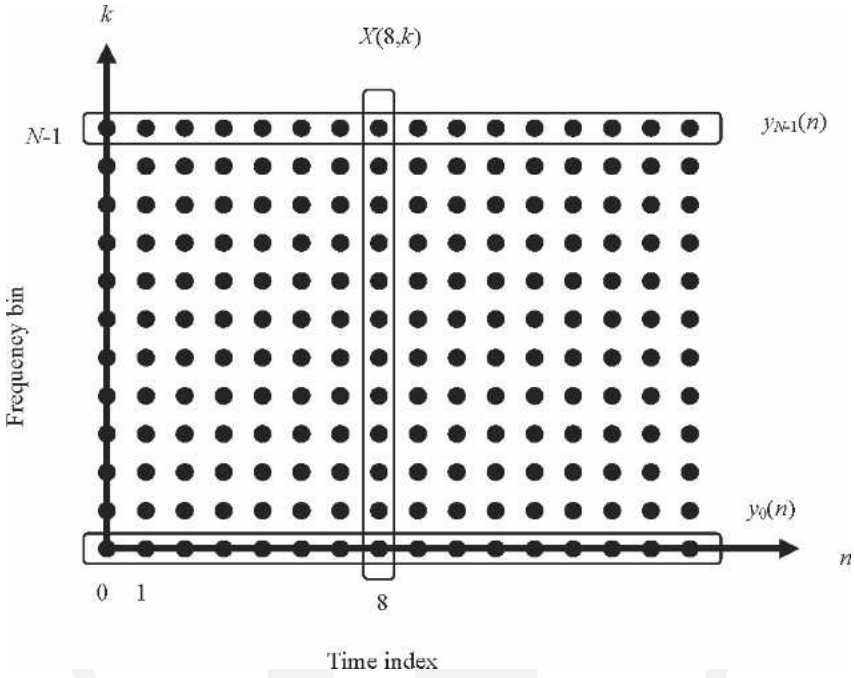


Fig. 2. STFT of signal with two-dimensional representation in phase-vocoder approach.

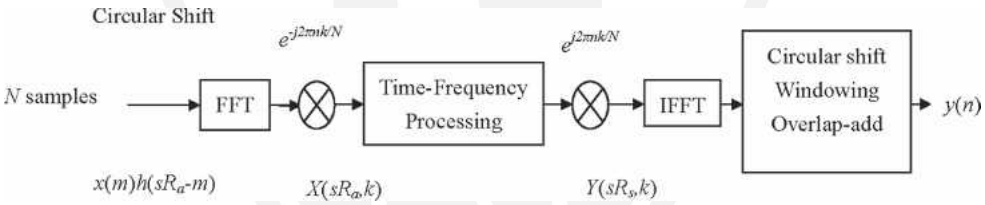


Fig. 3. FFT/IFFT structure for phase-vocoder implementation.

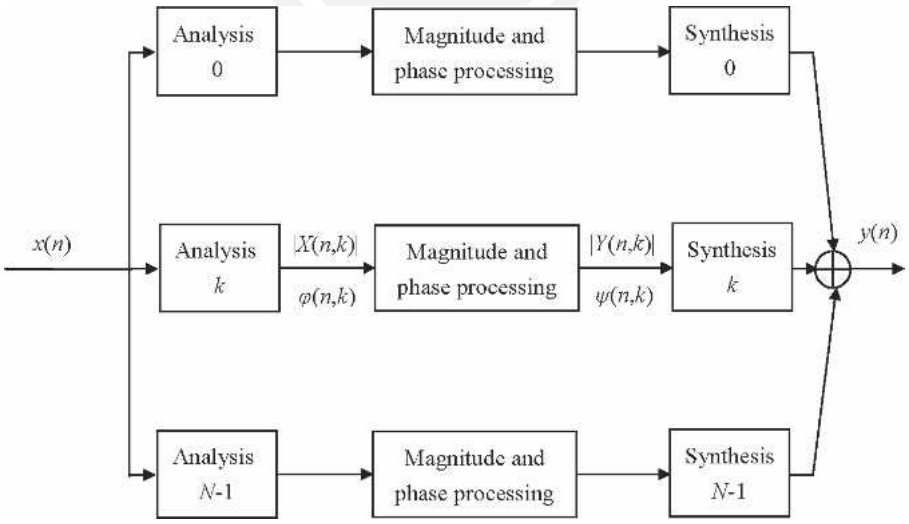


Fig. 4. FFT/sum-of-sinusoids structure for phase-vocoder implementation.

procedure is required. The unwrapping process corrects phase angles to produce continuous phase plots, which is crucial in calculating instantaneous frequency (phase increment per sample). Instantaneous frequency can be estimated from successive STFT. For a given value of k, computing the backward difference of the STFT phase yields

$$\Delta\phi = \phi[(s+1)R_s, k] - \phi(sR_s, k) + 2m\pi = R_s f_i + 2m\pi \tag{4}$$

in which the instantaneous frequency f_i is assumed to be constant over the hop size of R_a. The term $2m\pi$ stems from the fact that only the principal value of the phase is restricted to be within $(-\pi, \pi)$ during FFT. For a sinusoid to be resolved in the kth bin, the following condition must be satisfied:

$$|(\Omega_k - f_i)R_a| \leq \omega_h \tag{5}$$

where ω_h is the bandwidth of the analysis window. If R_a is such that $\omega_h R_a < \pi$, then

$$|\Delta\phi - \Omega_k R_a - 2m\pi| < \pi. \tag{6}$$

Once m is determined by this "unwrapping" process, the instantaneous frequency can be obtained by

$$f_i = \Omega_k + \frac{1}{R_a}(\Delta\phi - \Omega_k R_a - 2m\pi). \tag{7}$$

As a consequence, the instantaneous frequency is estimated while retaining the phase coherence.

1.3 Pitch Shifting

Pitch shifting is used as the key element to generate superharmonics for the VB system. In this section it will be demonstrated how pitch shifting can be accomplished by using the phase-vocoder approach.

It should be clear that pitch shifting is different from frequency shifting. By definition, frequency shifting applies a constant addition to every frequency component, whereas pitch shifting performs a constant multiplication with every frequency component. Pitch shifting alters the instantaneous frequency by a constant factor while maintaining phase coherence. The modified instantaneous frequency in effect changes the pitch of the signal after synthesis. The pitch-up effect is achieved using a pitch-shift parameter greater than unity, whereas the pitch-down effect occurs when using a pitch-shift parameter less than unity. Let $\phi(k)$ and $\psi(k)$ be the phases in the analysis and the synthesis stages, respectively, and α the prespecified pitch-shift parameter. The procedures of pitch shifting can be summarized as follows.

1) An increment of phase is obtained from two consecutive frames after FFT and divided by the analysis hop size R_a (identical in both the analysis and the synthesis stages) to estimate the instantaneous frequency of the original signal,

$$d\phi(k) = \Delta\phi(k)/R_a \tag{8}$$

where $\Delta\phi(k)$ is an unwrapped phase.

2) Multiply the instantaneous frequency by a pitch-shift parameter α and integrate it to obtain the modified phase increment per sample,

$$\psi(n+1, k) = \psi(n, k) + \alpha \cdot d\phi(k). \tag{9}$$

3) Synthesize the signal by summing the sinusoids with amplitudes unchanged,

$$y(n) = \sum_{k=0}^{N/2} A(n, k) \cos[\psi(n, k)]. \tag{10}$$

2 SYNTHESIS OF THE VIRTUAL BASS EFFECT

In this section the implementation of VB synthesis will be discussed. Ever since the invention of loudspeakers, there has been a longstanding interest in high acoustical output at low frequencies. This is especially true for modern consumer electronics such as flat TV, laptop computers, mobile phones, PDAs, and portable audio. For small loudspeakers this presents a difficult design problem because of the limited volume velocity that is achievable at low frequencies. The small diaphragm area, a small mass, and high stiffness of small loudspeakers lead to a high resonance frequency and hence a low response level at low frequencies. The traditional way of bass enhancement is by direct amplification of the bass portion. However, the performance of direct bass boosting is limited by the finite cone excursion and power-handling capacity of the loudspeaker. This method can only enhance frequencies slightly below the resonance frequency. Overdriving the loudspeaker could result in nonlinear distortion or even ultimate damage. Another less costly but effective solution is to use a virtual bass system, with the bass enhancement being part of the auditory system, instead of extending the actual physical bandwidth. This motivates the development of psychoacoustics-based bass enhancement systems. In a VB system the bass impression is created psychoacoustically without actually producing low-frequency signals. Such a system exploits the bandwidth extension property of human hearing, that is, humans are capable of "extrapolating" the missing fundamental in the low-frequency range based on higher harmonics [4]. MaxxBass represents one of the commercial systems that made use of this psychoacoustic property [2].

2.1 MaxxBass—A Benchmark Method

The essential element of the MaxxBass system is a nonlinear device that converts frequencies below the loudspeaker resonance frequency to a series of harmonically related high frequencies. The nonlinear device employed in the MaxxBass is a multiplier embedded in a feedback loop, as shown in Fig. 5. It can be verified easily that an infinite number of harmonics can be generated by such a feedback multiplier. A potential problem that could arise with this type of nonlinear processing is the artifacts due to intermodulation distortion [1]. Apart from a harmonic generator giving the correct pitch, a VB system generally

requires some mechanism to adjust the magnitude of each harmonic component so that the timbre (spectral envelope) and loudness are close to what would be perceived over an ideal loudspeaker. MaxxBass introduced a loudness analyzer to determine the weighting for each harmonic based on the SPL-to-phon expansion ratio $R(f)$ [3]. For example, if the sound pressure level (SPL) at the input is decreased by X dB, the loudness at the output will be decreased by RX phons. The SPL-to-phon expansion ratio is a function of the frequency f and can be written as

$$R(f) = \frac{1.0}{\ln(f) \cdot 0.241 - 0.579}. \quad (11)$$

On the other hand, the SPL-to-phon residue expansion ratio for the nth harmonic is given by

$$RR(f, n) = 1 - 1\text{n}(n) \cdot 0.241 \cdot R(f). \quad (12)$$

According to MaxxBass, the energy of the harmonic appropriate for the same loudness as the fundamental frequency can be approximated by

$$E_{\text{h}}(f, n) = RR(f, n) \cdot E_{\text{f}} + \rho \quad (\text{dB}) \quad (13)$$

where $E_{\text{h}}(f, n)$ is the energy of the nth harmonic in dB, E_{f} is the energy of the fundamental frequency component in dB, and ρ is a small number for fine adjustment of the energy.

2.2 Phase-Vocoder–Based VB System

In contrast to the time-domain methods such as MaxxBass, a frequency-domain method is used in the present work to implement the VB system. Unlike the nonlinear device in MaxxBass and the modulation scheme used in the system suggested in [4], the harmonic generator proposed in this paper is based on the aforementioned pitch-shifting phase vocoder. The low-frequency signals are shifted to the intended harmonic in the high-frequency range.

The pitch-shifting effect is used to scale the frequency f by a factor α so as to produce a new signal with higher frequency f',

$$f' = \alpha \cdot f. \quad (14)$$

As compared to the other methods based on time-domain nonlinear processing, the phase-vocoder approach has the advantage that better timbral quality can be achieved by preserving the phase coherence. In addition, artifacts may arise in using nonlinear processing if intermodulation distortions are not handled carefully. In Section 3 the phase-vocoder technique will be compared with a time-domain nonlinear-processing–based method, MaxxBass, in listening experiments.

The proposed VB system is depicted in the block diagram of Fig. 6. First a low-pass filter with the cutoff frequency of the loudspeaker is applied to the input signal. Second, the low-pass-filtered signal is down-sampled and windowed using a Hanning window and then transformed to the frequency domain using FFT. The FFT size is 42 ms, which is the same as the tap number of the Hanning window. Third, the phase-vocoder–based pitch-shifting algorithm is applied to generate the required harmonics at the high frequencies. It should be noted that the signals below cutoff frequency are divided into three bands, H_1, H_2, and H_3, in the frequency domain, as shown in Fig. 7. Every frequency component in a band needs to be shifted with the same ratio to the high frequency. For example, a band covering 80–120 Hz will be shifted to a new band covering 160–240 Hz to serve as the second harmonic, to another band covering 240–360 Hz to serve as the third harmonic, and so forth. In general, generating up to the fourth harmonic suffices. Next those revised spectra are converted to the time-domain waveform by utilizing the method of the sum-of-sinusoids structure. Finally, with a hop size of 512, the same procedure is repeated to process the next frame.

The harmonic signals generated must be weighted appropriately to reflect proper timbre and loudness. In the present work these weighs are determined psychoacoustically according to equal-loudness contours [6], as shown

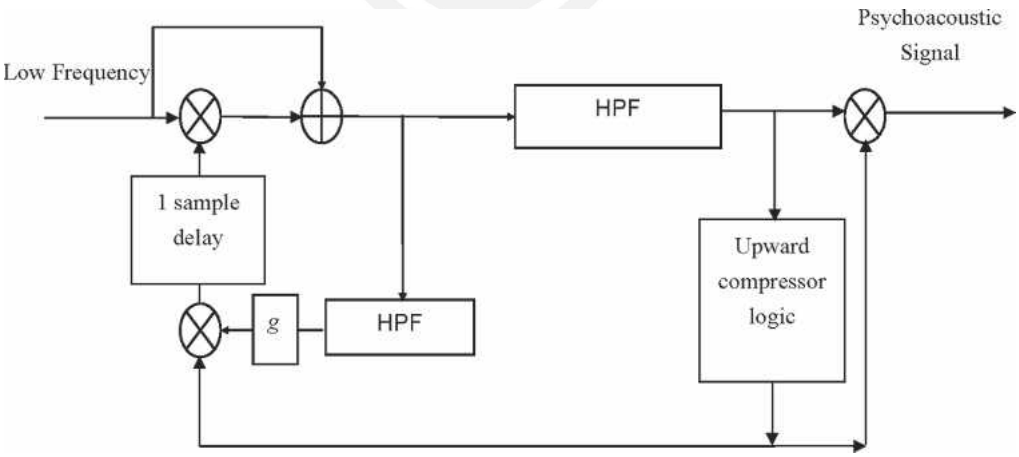


Fig. 5. Schematic diagram of MaxxBass system. Upward compressor logic in this loop is used to calculate SPL-to-phon expansion ratio and SPL-to-phon residue expansion ratio. g—constant gain; HPF—high-pass filter in feedback loop.

in Fig. 8. Close inspection of these contours reveals that, as the contours slope downward at low frequencies, human hearing is insensitive in terms of loudness at low frequencies, which must be taken into account in the design of the VB system. The contours are more compressed for very low frequencies than for higher frequencies. Consequently if we vary the level of two pure tones of unequal frequency by the same amount, then the loudness variation of the two will be unequal. The lower frequency tones will appear to have greater variation than the higher frequency tones. For the VB synthesis this would imply that if low-frequency components are replaced by high-frequency components, the loudness variations decrease. The signal shifted to the high frequencies should be properly attenuated to yield the same loudness as the low frequencies. The equal-loudness contours can be parameterized by the equation [6]

$$L_N = 4.2 + \frac{a_f(L_f - T_f)}{1 + b_f(L_f - T_f)} \quad (15)$$

where L_f and L_N denote SPL (dB) and loudness (phon), respectively, and a_f, b_f, and T_f are frequency-dependent parameters listed in Table 1. In the present VB system these frequency-independent parameters are further fitted into polynomials, as described by Eqs. (16)–(18),

$$a_f = -3.3378 \times 10^{-19} f^5 + 1.0889 \times 10^{-14} f^4 - 1.2776 \times 10^{-10} f^3 + 6.5607 \times 10^{-7} f^2 - 0.0014f + 1.8113 \quad (16)$$

$$b_f = -9.1993 \times 10^{-22} f^5 + 3.1889 \times 10^{-17} f^4 - 4.0519 \times 10^{-13} f^3 + 2.3588 \times 10^{-9} f^2 - 5.9306f + 0.0040 \quad (17)$$

$$T_f = 9.3706 \times 10^{-21} f^6 - 3.0490 \times 10^{-16} f^5 + 4.1801 \times 10^{-12} f^4 - 2.6922 \times 10^{-8} f^3 + 8.3228 \times 10^{-5} f^2 - 0.1115f + 46.48. \quad (18)$$

These polynomials are accurate in a frequency range of 2 kHz. This is a saving of memory storage for loudness contour parameters. Fig. 9 shows the close agreement between the data regenerated by the polynomials and the original data obtained using the parameter set a_f, b_f, and T_f. With these polynomials the parameters a_f, b_f, and T_f for arbitrary frequency can be obtained and the loudness can be calculated accurately and efficiently. Thus we can start with estimating the SPL of the fundamental and calculate

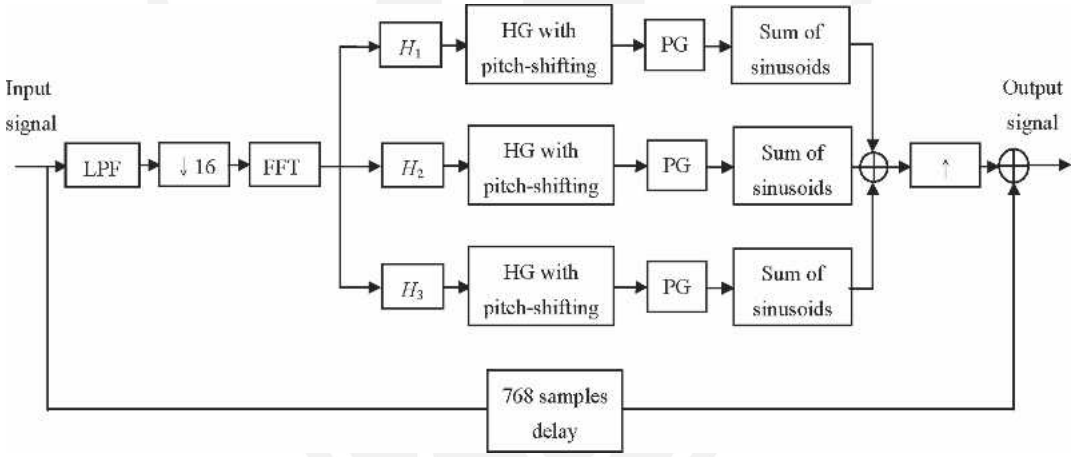


Fig. 6. Block diagram showing implementation of VB system. Up or down sampling ratio is 16; FFT size is 2048. Three bands H_1, H_2, and H_3 are equally spaced in frequency domain. HG—harmonic generator; PG—psychoacoustic gain control.

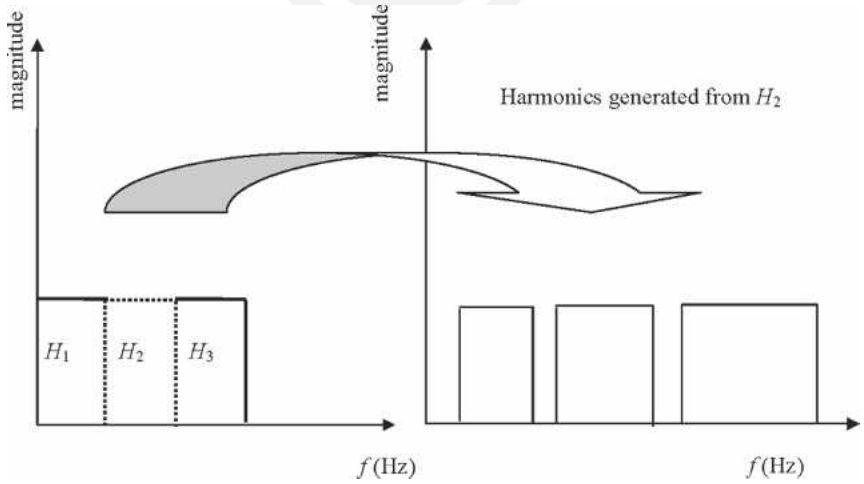


Fig. 7. Schematic diagram shows idea of harmonic generation of VB system. Only processing of band H_2 is shown.

the corresponding loudness, followed by the SPL of the harmonic components with the same loudness. After FFT, the energy of the fundamental components and the harmonics generated can be calculated by summing the magnitude squares in each band on a linear scale. Consequently the generated harmonics should be gain adjusted according to

$$w_{\mathrm{h}} = \frac{E'_{\mathrm{hVB}}}{E_{\mathrm{hVB}}} \tag{19}$$

where E_{hVB} is the square root of energy for a harmonic component with equal loudness as the target fundamental, and E'_{hVB} is the square root of energy for the gain-adjusted harmonic. According to Eq. (19) the generated harmonics are multiplied by a gain w_{h}. As a result the gain-adjusted harmonic yields the same loudness as the fundamental low frequency. The amount of attenuation needed for comparable loudness can then be determined from the difference between the two SPLs.

2.3 Implementation Issues

Since only very low frequencies, such as below 120 Hz, are of concern, a 16-fold down sampling is used in the VB system to reduce the original sampling rate of 48 kHz to 3 kHz. An additional benefit of this is that more processing time is available within one sample period under the low rate. Note that the fourth harmonic at 480 Hz is still within the Nyquist frequency of 1.5 kHz. After all harmonics are created, up sampling is required to restore the original sampling rate, 48 kHz. Multirate processing along with polyphase representation can be exploited to carry out up or down sampling in Fig. 6 quite efficiently [8].

Another caveat should be pointed out. In certain bands the second, third and fourth harmonic components could fall below the cutoff frequency, say, 120 Hz. For instance, a 40-Hz signal is supposed to generate the harmonics 80 Hz, 120 Hz, and 160 Hz, but the second harmonic falls below the cutoff frequency. In this case, the third, fourth, and fifth harmonics are generated using the VB system instead of the second, third, and fourth harmonics.

Table 1. Parameters of equal-loudness contours.

Frequency (Hz)	a_f (dB^{-1})	b_f (dB^{-1})	T_f (dB)
20	2.347	0.00561	74.3
25	2.190	0.00527	65.0
31.5	2.050	0.00481	56.3
40	1.879	0.00404	48.4
50	1.724	0.00338	41.7
63	1.597	0.00286	35.5
80	1.512	0.00259	29.8
100	1.466	0.00257	25.1
125	1.426	0.00256	20.7
160	1.394	0.00255	16.8
200	1.372	0.00254	13.8
250	1.344	0.00248	11.2
315	1.304	0.00229	8.9
400	1.256	0.00201	7.2
500	1.203	0.00162	6.0
630	1.135	0.00111	5.0
800	1.062	0.00052	4.4
1000	1	0	4.2
1250	0.967	−0.00039	3.7
1600	0.943	−0.00067	2.6
2000	0.932	−0.00092	1.0
2500	0.933	−0.00105	−1.2
3150	0.937	−0.00104	−3.6
4000	0.952	−0.00088	−3.9
5000	0.974	−0.00055	−1.1
6300	1.027	0	6.6
8000	1.135	0.00089	15.3
10000	1.266	0.00211	16.4
12500	1.501	0.00488	11.6

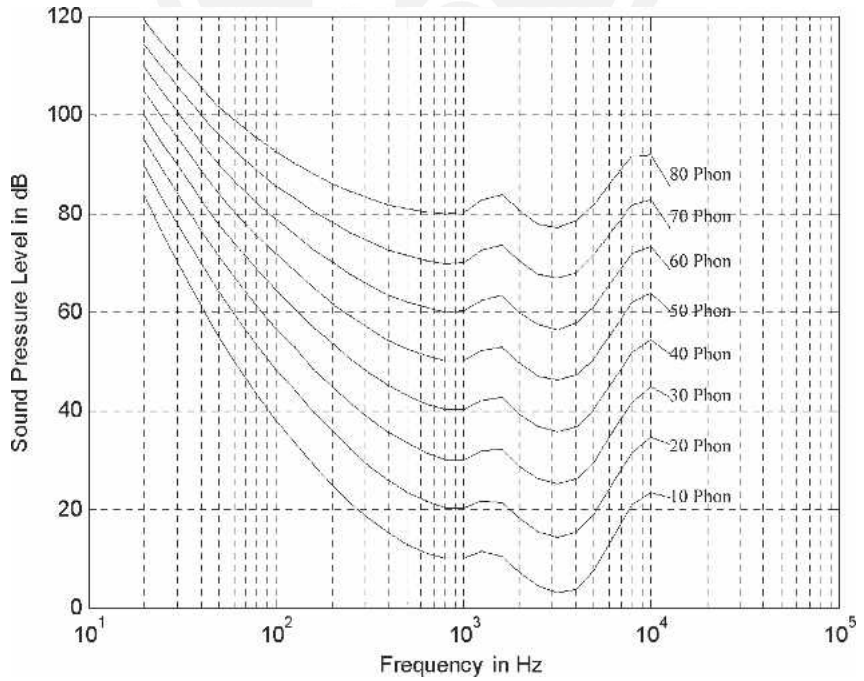


Fig. 8. Equal-loudness contours.

By using multirate up or down sampling and the polynomial model, the computational cost can be reduced substantially. Compared to the VB approach, MaxxBass requires 2235 instruction cycles per sample, while the VB only needs 445 cycles per sample.

3 EXPERIMENTAL INVESTIGATIONS

In order to validate the proposed VB system, objective and subjective tests were carried out.

3.1 Objective Experiments

As a preliminary test of pitch shifting using a phase vocoder a sinusoid at 1 kHz was shifted to generate the second, third, and fourth harmonics, as shown in Fig. 10. The harmonics were generated at precise frequencies, as intended. Fig. 11 shows the time-domain waveform and the associated spectrogram of an unprocessed pop music clip. Fig. 12 illustrates the pitch-shifted version of the same signal, where the pitch-shift parameter was selected

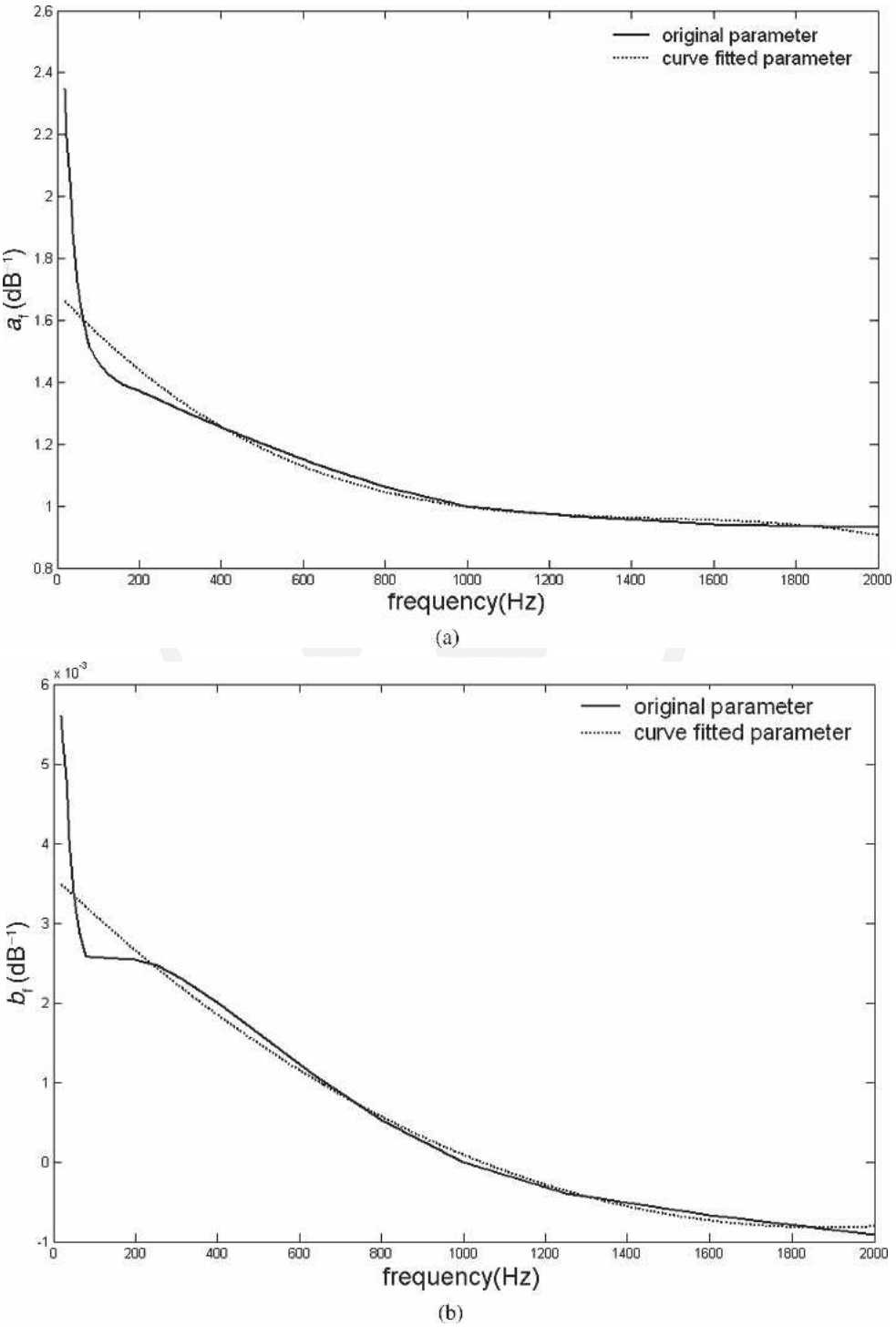
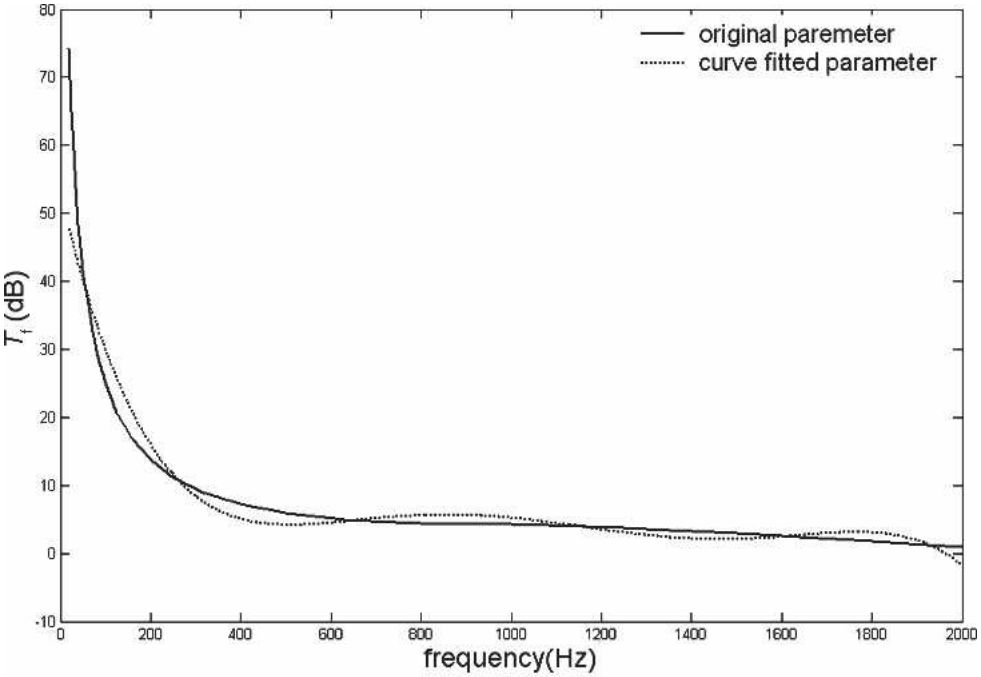


Fig. 9. Curve-fitted polynomial of parameters of equal-loudness contours. (a) Original and curve-fitted a_f. (b) Original and curve-fitted b_f. (c) Original and curve-fitted T_f.

to be 3. The results indicate that the frequency content of the original signal was moved to high frequencies with the duration unchanged. Next the proposed VB system was applied to process a pop music clip with a duration of 9.5 s. The transducers were the handset microspeakers. The FFT size was 2048. To facilitate the comparison, instead of time–frequency diagrams, spectra of the unprocessed signal and the VB-processed signal are shown in Fig. 13. As compared to the spectrum of the unprocessed signal in Fig. 13(a), a notable increase in level from 120 to 480 Hz can be seen in the VB-processed result in Fig. 13(b). The low-frequency components below 120 Hz are removed and replaced by synthesized harmonics at high frequencies. It is these generated harmonics that will psychoacoustically enhance the bass perception, as will be shown next in the listening experiments.



(c)

Fig. 9(c). *Continued*

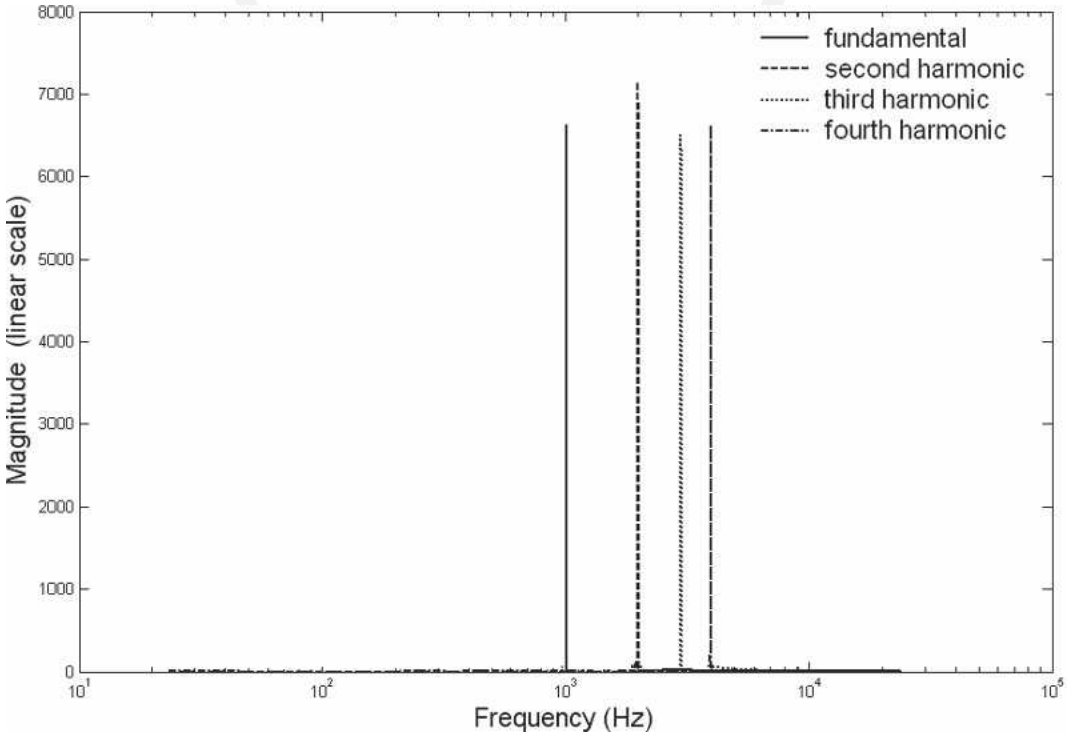


Fig. 10. Example of harmonics generated by pitch-shifting algorithm from a sinusoid at 1 kHz.

3.2 Subjective Listening Tests

Since the VB system proposed in this research is psychaoustics-based, it would be convincing to verify it with subjective listening tests. Sixteen subjects participated in these tests. The loudspeaker arrangement and listening room follows ITU-R BS.1116 [9]. Multistimuli with hidden reference and anchor (MUSHRA) of ITU-R BS.1534-1 [10] was used as the test procedure. Two audio attributes were selected to be the subjective indices for the tests:

1) *Bass impression* Dominance of low-frequency sound.

2) *Audio quality in terms of noise and distortion* Any extraneous disturbances to the signal are considered noise. Effects on the signal that produce new sound or timbre change are considered distortion.

The subjects participating in the tests were instructed with the definitions of the subjective indices prior to the tests. In the listening tests the subjects were asked to respond to a questionnaire with the subjective indices placed on a scale from 1 to 5. Scales from 1 to 5 indicate the qualities of bad, poor, fair, good, and excellent. The performances of the proposed phase-vocoder–based VB system and the MaxxBass system were compared. A pop music excerpt was used as the stimulus. In the MUSHRA test the unprocessed signal was used as the hidden reference and the high-pass-filtered signal as the anchor. The anchor does not contain frequency components below the loudspeaker cutoff frequency. The VB algorithms were implemented on a floating-point digital signal processor (DSP), ADI SHARC 21161. The test cases could be selected randomly by the subjects. Finally in order to justify the statistical significance of the test results, the scores

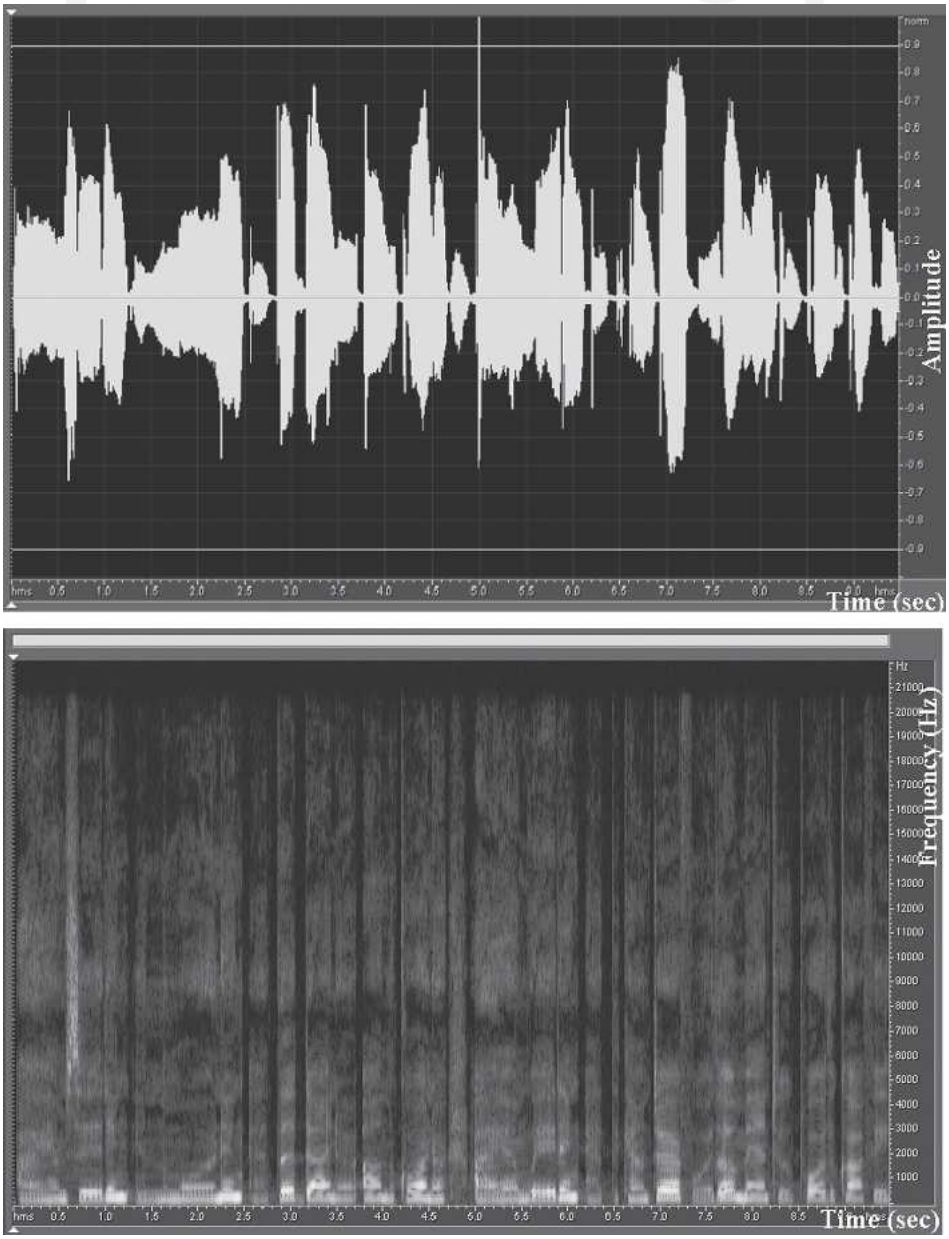


Fig. 11. Unprocessed signal (duration 9 s) displayed in time domain and time–frequency domain.

from all subjects were processed using the MANOVA [11]. Cases with p values below 0.05 indicate that the difference between cases is statistically significant.

In the first listening test a multimedia two-channel stereo loudspeaker with 120-Hz cutoff frequency was used as the rendering device, as shown in Fig. 14(a). Four stimuli obtained using the proposed VB system, the MaxxBass, the hidden reference, and the anchor were reproduced for the test. The grades of the bass impression and the audio quality are shown in Fig. 15, where the vertical bars indicate the 0.95 confidence intervals. The p value in the MANOVA output was only 0.00001, indicating that the difference among the methods is statistically significant. As expected, the anchor attained the lowest performance in bass impression because its low-frequency content has been filtered out. Both the MaxxBass system and the VB system produced better bass impression than the unprocessed reference. However, the proposed VB system significantly outperformed the MaxxBass system by nearly one point in the test. On the other hand, in terms of audio quality the stimuli processed using the VB system and the MaxxBass system did not seem to perform as well as the unprocessed reference because of some unnatural sensation reported by some of the subjects. Nevertheless, most listeners still considered that the proposed VB system produced less distortion than the MaxxBass system. Overall the VB system presented in this paper is superior to the MaxxBass system in enhancing bass performance of the multimedia loudspeakers.

In the second listening test we wished to explore what the subjective perception of bass would be in the system played on small multimedia loudspeakers, as compared to the original signal played on an equivalent larger loudspeaker. The experimental arrangement of this test was the

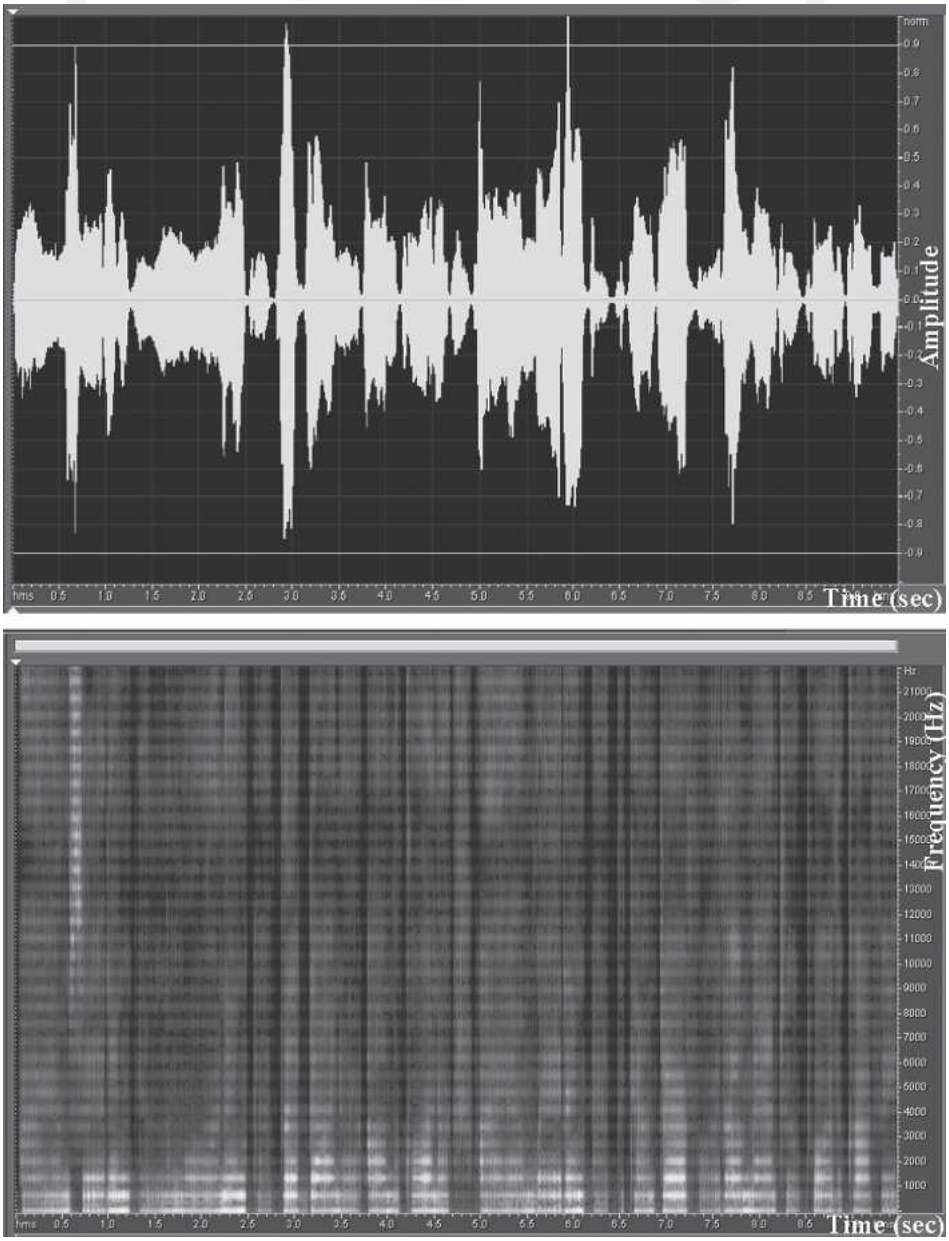


Fig. 12. Pitch-shifted signal (duration 9 s) displayed in time domain and time–frequency domain with pitch-shifting parameter = 3.

same as in Fig. 14(a). The VB-processed stimuli were reproduced with a 3-in (76-mm) multimedia loudspeaker, whereas the unprocessed signal was reproduced with a 5-in (127-mm) loudspeaker ($f_0 = 60$ Hz), which served as the hidden reference. The anchor in this test was the high pass-filtered signal in which the input below the cutoff frequency ($f_0 = 120$ Hz) of the 3-in (76-mm) loudspeaker is removed. The stimuli were compared in the listening test in terms of bass impression and audio quality. The results of the test and the associated ANOVA analysis are shown in Fig. 16. The p value in the ANOVA output was 0.00843, meaning that the differences between these stimuli were statistically significant. As expected, the anchor attained the worst performance in bass impression. The bass impression created by the VB system reproduced using a 3-in (76-mm) loudspeaker was equivalent to that created by the unprocessed signal reproduced using a 5-in (127-mm) loudspeaker.

As a more difficult problem, a handset equipped with dual microspeakers (diameter 16 mm, cutoff frequency 800 Hz) was used in the next experiment. The experimental arrangement of this test is shown in Fig. 14(b). The stimuli were the proposed VB system, the MaxxBass, and the hidden reference, as defined previously. Somewhat

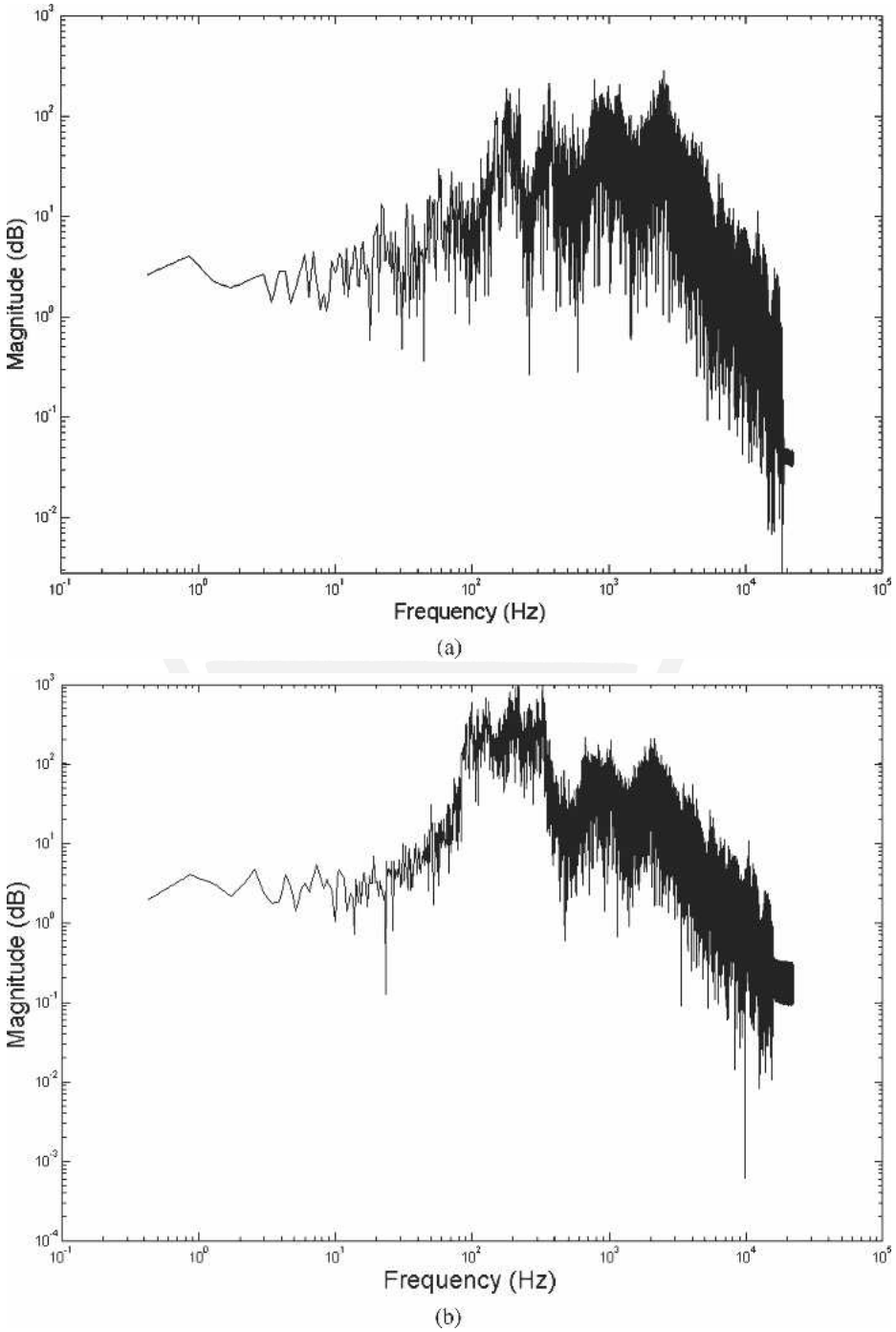


Fig. 13. Spectrum of 46-ms pop music clip obtained from VB system. Transducers are handset microspeakers. FFT size 2048. (a) Unprocessed signal. (b) VB-processed signal.

different from the previous listening test, the anchor in this test was a high-pass-filtered signal in which the input below the cutoff frequency of the microspeakers was removed. The results of the listening test in terms of bass impression and audio quality are shown in Fig. 17. The p value in the MANOVA output was only 0.00213, indicating that a significant difference exists between the stimuli. As expected, the anchor attained the worst performance in bass impression. The proposed VB system produced better bass impression than the unprocessed reference. Similar to the outcome of the multimedia loudspeaker, the VB system appeared to be slightly inferior to the unprocessed reference in terms of audio quality. Nevertheless, most listeners still considered the somewhat artificial sensation produced by the VB system acceptable. In particular, the VB system delivered significantly better performance in both audio attributes than the benchmark method MaxxBass. Thus it is fair to say that the present VB system is effective in enhancing the bass performance of the microspeakers of handsets.

4 CONCLUSIONS

A psychoacoustic bass enhancement technique for loudspeakers has been developed. Unlike the other psychoacoustic VB methods that rely on nonlinear processing, the proposed technique is based on a phase-vocoder approach to generate the required harmonics at high frequencies. The virtue of pitch shifting using phase vocoder is that phase coherence is well preserved, while traditional nonlinear processing may produce undesirable intermodulation distortion. The present VB system estimates the low-frequency SPL and determines the proper weights for each generated harmonic in light of a polynomial model of equal-

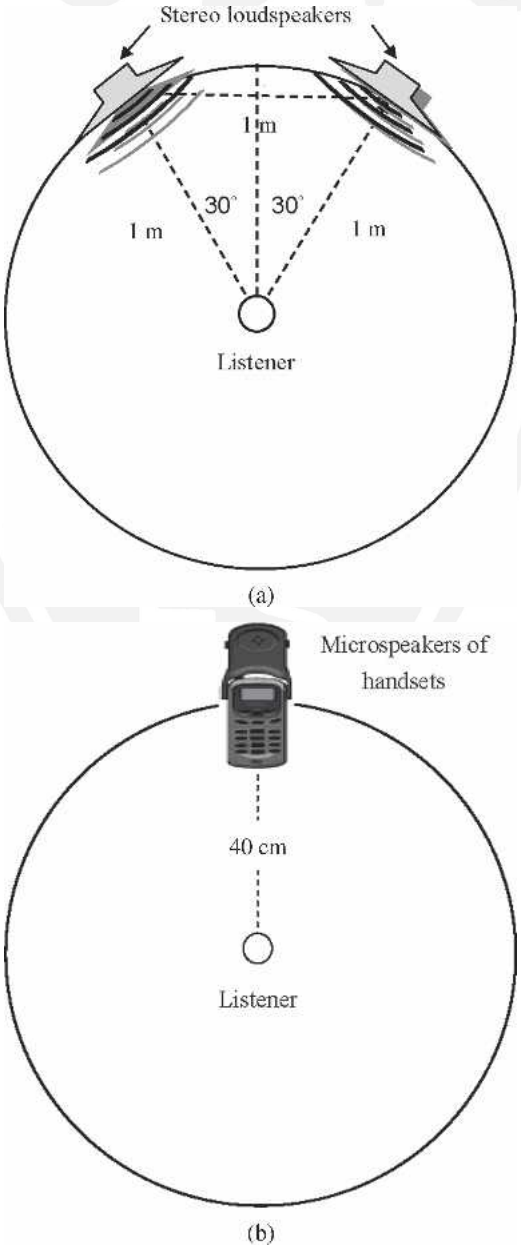


Fig. 14. Experimental arrangements of subjective listening test. (a) Multimedia stereo loudspeakers. (b) Handset microspeakers.

loudness contours. Up or down sampling is used to improve the computational efficiency of the VB synthesis. The VB system has been implemented on a floating-point DSP.

Subjective listening tests were conducted to assess the performance of the presented VB system versus a benchmark MaxxBass system, with application in multimedia loudspeakers and handset microspeakers. The subjective listening experiment followed the MUSHRA procedure and the data were analyzed using the MANOVA method. The statistical analysis revealed that the proposed VB approach was capable of rendering a superb bass impression with acceptable audio quality.

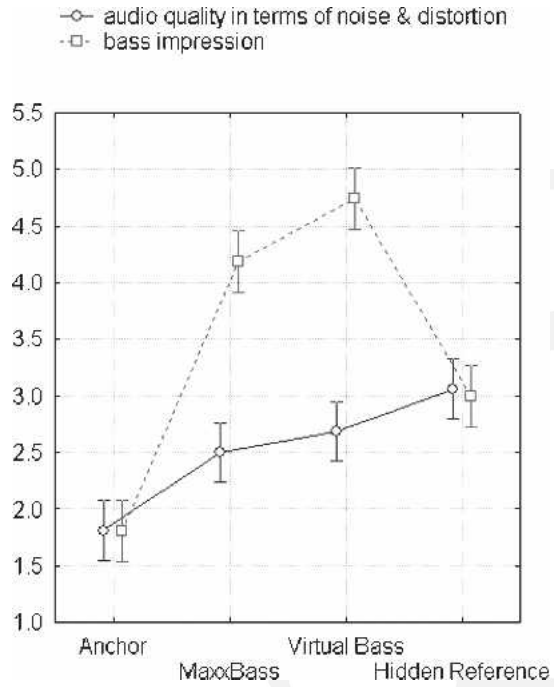


Fig. 15. Grades of listening test (mean and spread with 95% confidence interval) for multimedia stereo loudspeakers. p value = 0.00001.

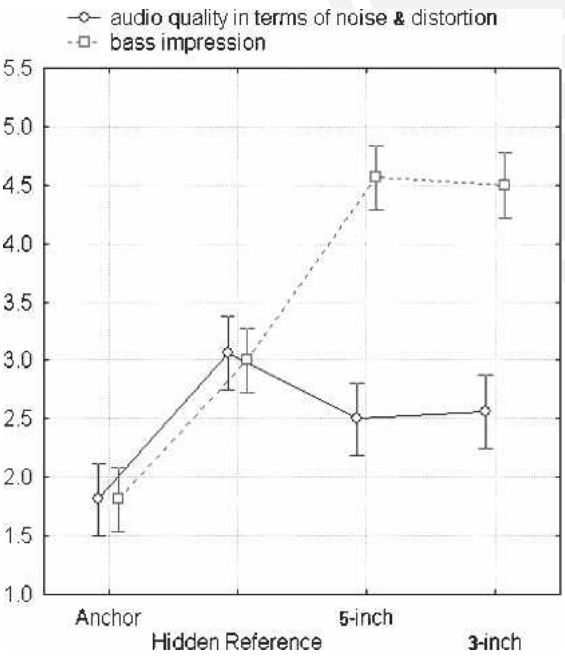


Fig. 16. Grades of listening test (mean and spread with 95% confidence interval) is for two different size multimedia stereo loudspeakers. p value = 0.00843.

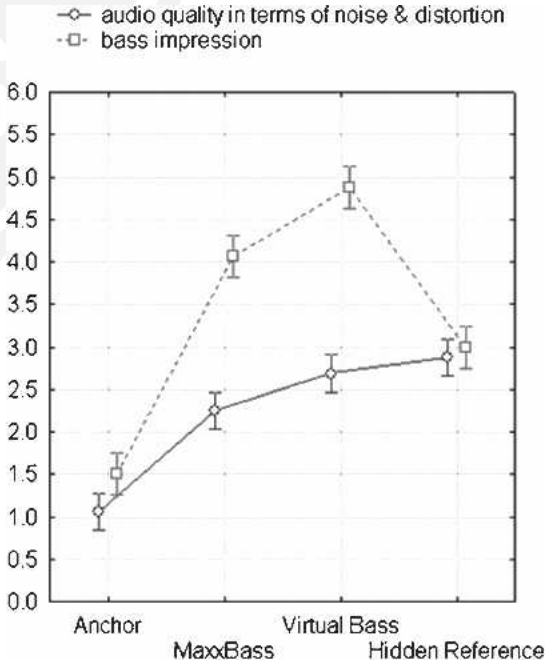


Fig. 17. Grades of listening test (mean and spread with 95% confidence interval) for handset microspeakers. p value = 0.00213.

5 ACKNOWLEDGMENT

The work was supported by the National Science Council in Taiwan, Republic of China, under the project NSC94-2212-E-009-019.

6 REFERENCES

[1] E. Larsen and R. M. Aarts, *Audio Bandwidth Expansion* (Wiley, West Sussex, UK, 2004).

[2] B. T. Daniel and C. Martin, "The Effect of the MaxxBass Psychoacoustic Bass Enhancement System on Loudspeaker Design," http://www.maxx.com/ (2000).

[3] M. Shashoua and D. Glotter, "Method and System for Enhancing Quality of Sound Signal," US Patent 5930373 (1999).

[4] W. S. Gan, S. M. Kuo, and C. W. Toh, "Virtual Bass for Home Entertainment, Multimedia PC, Game Station and Portable Systems," *IEEE Trans. Consumer Electron.,* vol. 47, pp. 787–794 (2001).

[5] J. Laroche and M. Dolson, "New Phase-Vocoder Techniques for Real-Time Pitch Shifting, Chorusing, Harmonizing, and Other Exotic Audio Modifications," *J. Audio Eng. Soc.,* vol. 47, pp. 928–936 (1999 Nov.).

[6] D. W. Robinson and R. S. Dadson, "A Redetermination of the Equal-Loudness Relations for Pure Tones," *Brit. J. Appl. Phys.,* vol. 7, pp. 166–181 (1956).

[7] R. E. Crochiere and L. R. Rabiner, *Multirate Digital Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1983).

[8] P. P. Vaidyanathan, *Multirate Systems and Filter Banks* (Prentice-Hall, Englewood Cliffs, NJ, 1993).

[9] ITU-R BS.1116, "Methods for the Subjective Assessment of Small Impairments in Audio System Including Multichannel Sound Systems," International Telecommunication Union, Geneva, Switzerland (1994).

[10] ITU-R BS.1534-1, "Method for the Subjective Assessment of Intermediate Sound Quality (MUSHRA)," International Telecommunications Union, Geneva, Switzerland (2001).

[11] G. Keppel and S. Zedeck, *Data Analysis for Research Designs* (Freeman, New York, 1989).

[12] A. V. Oppenheim and R. W. Schafer, *Discrete-Time Signal Processing* (Prentice-Hall, Englewood Cliffs, NJ, 1999).

[13] U. Zolzer, Ed., *DAFX: Digital Audio Effects* (Wiley, New York, 2002).

[14] M. Kahrs, *Applications of Digital Signal Processing to Audio and Acoustics* (Kluwer Academic, New York, 1998).

[15] T. H. Andersen and K. Jensen, "Importance and Representation of Phase in the Sinusoidal Model," *J. Audio Eng. Soc.,* vol. 52, pp. 1157–1169 (2004 Nov.).

[16] J. Laroche and M. Dolson, "Improved Phase Vocoder Time-Scale Modification of Audio," *IEEE Trans. Speech Audio Process.,* vol. 7, pp. 323–332 (1999).

THE AUTHORS

M. R. Bai

W. Lin

Mingsian R. Bai was born in Taipei, Taiwan, Republic of China, in 1959. He received a bachelor's degree in power mechanical engineering from the National Tsing-Hwa University in 1981 and a master's degree in business management from the National Chen-Chi University in 1984. He left Taiwan in 1984 to enter the graduate school of Iowa State University and received an M.S. degree in mechanical engineering in 1985 and a Ph.D. degree in engineering mechanics and aerospace engineering in 1989.

In 1989 Dr. Bai joined the Department of Mechanical Engineering of the National Chiao-Tung University in Taiwan as an associate professor and became a professor in 1996. He was also a visiting scholar to the Center of Vibration and Acoustics, Pennsylvania State University, the University of Adelaide, Australia, and the Institute of Sound and Vibration Research (ISVR), United Kingdom, in 1997, 2000, 2002, respectively. His current interests encompass acoustics, audio signal processing, electroacoustic transducers, vibroacoustic diagnostics, and active noise and vibration control. He currently serves as an active consultant and project leader in these areas of the industry.

Dr. Bai has published over 100 papers and has 13 granted or pending patents. He is a member of the Audio Engineering Society, the Acoustical Society of America, the Acoustical Society of Taiwan, and the Vibration and Noise Control Engineering Society in Taiwan.

●

Wan-Chi Lin was born in Taipei, Taiwan, ROC, in 1981. She received a bachelor's degree and a master's degree in mechanical engineering from the National Chiao-Tung University in 2004 and 2006, respectively. Her master's thesis is on audio signal processing. She is currently an acoustic engineer with Chi Mei Communication Systems, Inc.