COMPARISON OF REAL-TIME VIRTUAL BASS ENHANCEMENT TECHNIQUES

Armando Boemio, Gabriele Maucione

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano
Piazza Leonardo Da Vinci 32, 20122 Milano, Italy
[armando.boemio,gabriele.maucione]@mail.polimi.it

ABSTRACT

In the recent years, every component in the consumer electronics has gone through a miniaturization process. This also includes loudspeakers in mobile devices and laptops. However, reduced form factor loudspeakers are not capable to reproduce low frequencies. Virtual bass enhancement (VBE) systems aim at increasing the perception of low frequencies by extending their bandwidth, exploiting the missing fundamental psychoacoustic phenomena. Both off-line and real-time solutions exist in literature, with the second ones being of more interest in the consumer electronics field of use because of their flexibility. This paper presents and compares different real-time low-frequency bandwidth extension approaches using time, frequency and hybrid time-frequency domain techniques. The results of the comparison are presented both in terms of tonotopic curves and MUSHRA subjective listening tests.

***Index Terms*—** Audio enhancement, bandwidth extension, real-time techniques

1. INTRODUCTION

Small size loudspeakers have a limited frequency range in which they can efficiently reproduce audio signals. Indeed, limited dimensions and cost constraints cause them to have poor radiation performances on the lower portion of the audible spectrum. The most significant example is the piezoelectric loudspeaker, which is also the most diffused type of transductor in thin-profile consumer electronics. The frequency response of an ideal piezoelectric device is approximately constant above the resonance frequency f_c, while it is proportional to f^2 below. The value of f_c depends on many factors, but it is usually around 1 kHz. This behaviour makes the sound radiation performance in the lowest frequency bandwidths very poor. For this reason, it has been a challenge to find a solution to increase the perception of low frequency musical sounds, such as kick drums and bass guitars.

At first, the use of audio equalizers to physically boost the energy in that portion of the spectrum might seem a reasonable choice. However, this approach eventually leads to significant distortion or even irreversible damages to the device. For this reason, it is necessary to approach the problem in a different way.

Virtual bass enhancement (VBE) systems allow to increase the perception of bass frequencies without physically boosting their energy. This is possible by exploiting the psychoacoustic effect known as *missing fundamental* or *virtual pitch*. According to this theory, it is possible to trick the hearing system to perceive a fundamental only having its harmonics. Therefore, VBE systems generate new harmonic components in the lower-mid frequency range, that is reproduced way more efficiently by the speaker. As extensively described by *Larsen and Aarts* [1], psychoacoustic bandwidth extension algorithms for low frequencies result into shifting the energy content from a lower portion of the spectrum to a higher portion.

Being a psychoacoustic phenomena, it involves the subjective perception of the sound, in terms of bass enhancement and overall quality. Many parameters can influence the overall quality of the perceived sound, such as the number of harmonic introduced and the relation among their amplitudes. Different VBE approaches allow to have control over those parameter in many ways. Generally, harmonics are weighted according to a matching criteria, based on either loudness or timbre. A general scheme for a VBE system is presented in Fig.1. For the stated reasons, the main block that characterizes it is the harmonic generator. Depending on how this block works, it is possible to identify two different typologies of VBE systems: time domain and frequency domain systems.

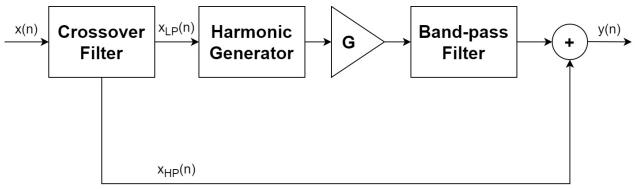

Figure 1: General architecture of the presented VBE systems.

Time domain techniques were the first to be studied and implemented. They rely on non-linear devices (NLD) to generate harmonics. Their main benefit is computational efficiency and flexibility, as it is possible to use several types of NLD, every of which has different desirable features. However, those techniques do not allow to directly control the number and the magnitude of the generated harmonics. Moreover, if the input signal is not a pure sinusoid, problems related to intermodulation distortion arise, that inevitably compromise both the timbre and the overall quality of the bass enhanced signal.

Frequency domain techniques were introduced later to overcome those drawbacks. The core of the harmonics generation is the phase vocoder (PV). This approach allows to work directly on the frequency spectrum to generate the desired number of harmonics with the desired amplitude. The drawbacks are a higher computational burden, and limitations due to frequency resolution, especially in a real-time environment. Also, phasiness artifacts and transient smearing problems may be present due to the PV structure.

Further steps into the research have been done. Hill *et al.* [2] proposed an hybrid approach to harmonic generation , trying to combine the best aspects of both techniques. This requires a transient/steady-state (TSS) separation stage, so that transient components are processed with time-domain techniques, while steady-

state components are processed with frequency domain techniques.

The paper presents and accurately compares three different VBE algorithms (one for each described technique) in a real-time application. Therefore, it is organized as follows. Sec. 2 describes the implementation of the algorithms, highlighting the advantages and the limits of each. Sec. 3 discusses the results in terms of tonotopic curves and MUSHRA subjective listening tests. Sec 4. concludes the paper.

2. PROPOSED REAL-TIME VBE ALGORITHMS

The general structure is common among all the proposed algorithms. The processing chain starts with a two-band 24 dB/octave crossover filter, whose crossover frequency f_c is 180 Hz. Since loudspeakers have different frequency responses and musical signals feature different bass ranges , the crossover frequency is tunable in between 130 and 250 Hz. Then, the mono low-passed signal is fed into the harmonic generator and weighting stage, that change according to the approach. After that, the enhanced bass signal is band-passed in the range 120-800 Hz with a 24 dB/octave filter. The high cut-off is also tunable, in order to have further control on the distortion introduced by the generated harmonics. The final stereo output is obtained by adding the enhanced signal to the high-passed stereo components.

2.1. Time Domain VBE

The choice of the non-linear device to use depends on many factors. Larsen and Aarts discussed and analyzed in [1] their differences extensively, in terms of spectral and time characteristics and intermodulation distortion.

In our time domain virtual bass system, the NLD of choice is a full-wave rectifier. Its implementation is trivial and therefore very inexpensive and suitable for a real-time application, while having at the same time very desirable temporal characteristics, as the temporal envelope remains almost identical to the original. The rectifier is also an amplitude linear device, therefore a very simple gain level adjust stage was implemented, so that the level of enhancement can be adjusted according to input signal and also subjective preference.

On the other hand, rectifiers only generate even harmonics, so if the fundamental is missing, it is possible to perceive an effect of pitch doubling. This may happen either if the fundamental is filtered or if the speaker is not able to reproduce it. Moreover, because of the rectifier, the algorithm presents poor or mediocre characteristics in terms of intermodulation distortion, depending on the input signal.

2.2. Frequency Domain VBE

Frequency domain algorithms allow to overcome the problem of intermodulation distortion, using a Phase Vocoder as harmonic generation block.

The proposed PV is based on a FFT/IFFT approach as shown in Fig. 2. After the transformation in the frequency domain, the algorithm estimates the fundamental frequency, generates its harmonics and weights them according to a timbre matching criteria. Finally, the enhanced signal is reconstructed and transformed back into the time domain.

Prior to any processing, the low-passed signal frame is downsampled by a factor 16 to reduce the FFT computational burden. The fundamental f_0 is estimated through a simple peak-finding function, as it is a reasonable choice to consider the frequency of the highest peak as the fundamental. For increased precision, the bin of the fundamental is estimated with parabolic interpolation.

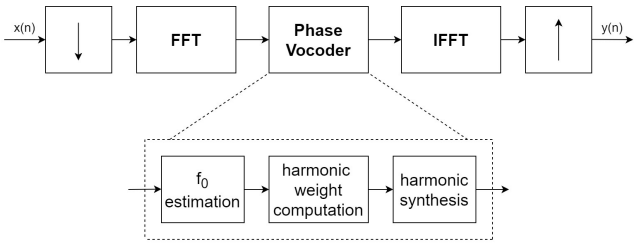

Figure 2: VBE harmonic generation block in the frequency domain.

The harmonics are generated according to the idea introduced in [3] by Laroche and Dolson. First, a region of influence (ROI) is defined around the fundamental bin. Then, the whole region is shifted by an integer multiple of the fundamental

$$f_k = k f_0 \tag{1}$$

where k is the harmonic order. Phase coherence in audio signal has a significant influence on the overall quality perceived. One of the used strategies, known as "Vertical Phase Locking" [4], is proved to guarantee high quality synthesized signals. Moreover, a "Horizontal Phase Locking" mechanism is also necessary to maintain synthesis phase coherence across time frames and avoid artifacts in the processed signal. This is done by satisfying the condition

$$\angle \mathrm{Y_{PV}}(m,n) = k \angle \mathrm{X_{PV}}(m,n) \tag{2}$$

as proven in [5]. In the proposed solution, four harmonics are generated, since it is proven to be enough for the auditory system to perceive a bass enhancement.

The last step is the weighting stage, realized according to a timbre matching criteria presented by Moliner *et al* in [6]. The spectral envelope of each frame is computed using a Bark scale triangular filter bank to the signal. The algorithm computes then two constant exponentially decaying curves, that define the lower and upper limits of the output envelope containing the generated harmonics. For further control over the enhancement effect, the user is given a tunable gain parameter to either boost or reduce the amplitudes of the harmonics.

Since the algorithm works in real-time, the frequency resolution of the FFT is limited but still sufficient to process small bandwidth downsampled signals. Moreover, being a frequency domain approach, also our algorithm is subject to occasional phasiness artifacts and transient smearing effects.

2.3. Hybrid VBE

In order to overcome both intermodulation distortion problems of time domain techniques and transient smearing due to the frame-based approach of STFT, a hybrid time-frequency sistem is proposed.

The main idea is to process steady state signals in the frequency domain and transients in the time domain. This requires a transient/steady-state (TSS) detector. In our algorithm it is implemented by a peak amplitude difference ODF (Onset Detection Function). This is based on the idea that in a frame based approach,

the harmonic components can be seen a sum of slowly varying sinusoid, therefore there is no significant change in both peak amplitudes and frequencies. Instead, transient regions at note onset locations should show considerably more frame-by-frame variation in both peak frequency and amplitude values [7]. The peak amplitude difference $\mathrm{ODF_{PAD}}$ is given by

$$\mathrm{ODF_{PAD}}(n) = \sum_{k=0}^{M_p} |P_k(n) - P_k(n-1)| \quad (3)$$

where *n* is the frame index, $\mathrm{P_k}(n)$ is the peak amplitude of the k-th partial. The algorithm works as follows:

Data: ODF
Result: Onset
Onset = false;
if $\mathrm{ODF}(n-1) \geq \mathrm{ODF}(n)$ & $\mathrm{ODF}(n-1) \geq \mathrm{ODF}(n-2)$ **then**
  if $\mathrm{ODF}(n-1) \geq \sigma_{TH}$ **then**
    Onset = true;
  end
end
UpdateValues();
return Onset;

Algorithm 1: Onset detection pseudo-code.

The algorithm presents a delay equal to 1 buffer, however this does not compromise the overall performance of the algorithm. The threshold value σ_{TH} depends on the weighted mean and median of the previous eight values of the ODF function. "Onset" is the flag that is raised if a transient is detected and determines the type of processing used. If it is true the input is processed in time domain, otherwise the frequency domain technique is used. Moreover, if no peaks over a certain threshold are found in the current frame, the signal is sent unprocessed to the output, so that false peaks due to noisy frames are not considered.

This technique is an improvement over the traditional PV approach, however it is possibe to hear some artifacts if too many onset events are detected in a small time-window. This causes the algorithm to frequently switch from one approach to another, that intrinsically present slight differences in the timbre or volume of the generated harmonics.

3. EXPERIMENTAL RESULTS

Virtual bass enhancement systems are based on a psychoacoustic effect as low frequencies are not physically enhanced in the radiated signal. The enhancement happens in a psychoacoustic fashion directly in the auditory system of the listener. Therefore, it is not possible to objectively evaluate the performances of such systems by looking at their frequency response. The most used tools to overcome this limit are tonotopic curves and subjective listening tests. The proposed algorithms are evaluated by means of both methods.

3.1. Tonotopic Curves

Tonotopic curves are a powerful tool when it comes to evaluate the auditory system response to an audio signal. They describe the excitation pattern of the basilar membrane, therefore it is possible to visualize how low frequency are perceived before and after the enhancement. The Auditory Image Model (AIM) was used to plot the Neural Activity Pattern for both the unprocessed example track and its processed versions, as shown in Fig. 3.

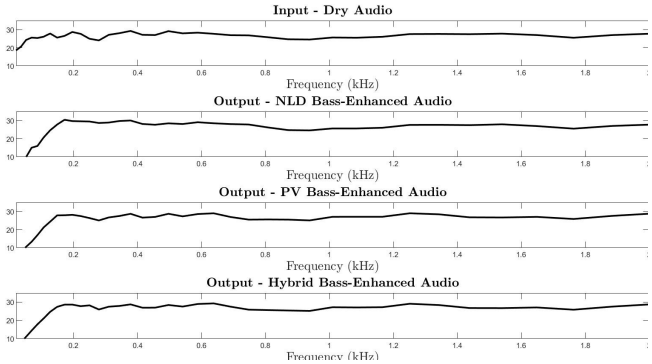

Figure 3: Tonotopic curves of a 200ms segment of an example track. From top to bottom: original signal, time-domain VBE, frequency-domain VBE, hybrid VBE.

It is possible to notice how the perception of low frequencies is enhanced around the fundamental, while the high-pass still causes a light dampening effect on sub-bass frequencies. However, this does not reflect into a significant loss of bass enhancement perception when using a small loudspeaker.

3.2. Subjective Listening Test

The MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) [8] was chosen to blindly compare the performance of the three presented algorithms. The proposed test presented 15 seconds excerpts of four different songs from different musical genres. The average duration of a test is less than 10 minutes, therefore it was not tiresome for the listener.

The songs were chosen to test the performances over different bass/kick drums patterns, tones and timbres. The songs are listed in Table 1. In particular, the funk track has rich mid-bass tones with a groovy rhythmic pattern; the hip-hop track is characterized by a predominant kick drum pattern; the jazz track was chosen for the peculiar timbre of the double-bass; the rock track is characterized by long electric bass notes.

The parameters of the algorithms are supposed to be fine-tuned to obtain the best quality/enhancement trade-off under different listening condition for each audio input. However, in order to keep the comparison as fair as possible, it was decided to assign an average value of gain and cut-off frequencies for all the algorithms and audio signal under test.

#	Title	Artist	Genre
1	Cloud 9	Jamiroquai	Funk
2	In da Club	50Cent	Hip-hop
3	L-O-V-E	Nat King Cole	Jazz
4	Learn to Fly	Foo Fighters	Rock

Table 1: List of song fragment used for the listening test.

For each song, the listener is presented with an original unprocessed sample to use as a reference. Then, five different versions of the reference are blindly proposed. For each version, it was asked to vote in a scale from 1 (very poor) to 10 (excellent) two different

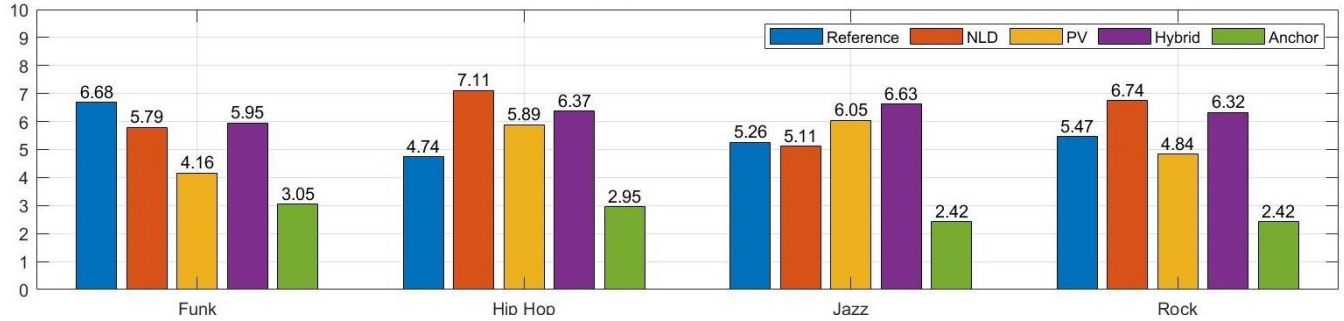

Figure 4: Mean scores of MUSHRA test for all the audio samples and conditions.

parameters: bass enhancement perception and overall quality. The 5 proposed versions included the three proposed algorithms, one hidden reference and one anchor, which is an high-passed version of the reference. The high-pass cut-off was set at 650 Hz with a -6dB/octave slope, in order to have a significant effect on the audio signal.

The subjects that completed the survey are a mix of expert and non-expert listeners. Some of them already had experiences with MUSHRA tests, others were not familiar with audio listening tests. Therefore the results may be slightly biased by non-optimal listening conditions and/or inexperience. Moreover, listeners were asked to complete the test on a personal mobile device or laptop, therefore speaker quality had a great influence over the bass enhancement evaluation. To account for this, the tests in which the listener was not able to recognize the anchor at least three over four times, were discarded. In particular, out of 27 subjects that completed the survey, 11 were discarded. The results are presented in Fig. 4 and table 2

	Bass Enhancement	**Overall Quality**
Reference	5.5375	7.0800
NLD	6.1875	6.4125
PV	5.2350	6.1825
Hybrid	6.3175	5.4350
Anchor	2.7100	4.6575

Table 2: Average performance ratings.

As expected, there is not an algorithm that works better in every condition. Phase Vocoder obtained worse ratings among the algorithms in terms of bass enhancement, while the Hybrid approach received the highest rating. In terms of overall quality, the reference was still preferred over the processed tracks, but the NLD approach was close second. In particular, averaging the two scores, the NLD is the most liked solution. The reference was rated higher than the processed tracks only for the funk genre. This can be explained by the presence of mid-bass tones, that may not be enhanced enough by the algorithm. Also, the complexity of the rhythmic pattern makes the phase vocoder inadequate for its intrinsic weakness on transients. NLD received quite poor results on the acoustic tones of the jazz track, but got the highest rating on hip-hop, because of a majority of transient kick sounds. The least favourite algorithm in terms of overall quality was the hybrid approach. This can be justified by the presence of artifacts due to the switching between time and frequency systems. Nonetheless, it showed to be the most robust and reliable system among the compared ones.

4. CONCLUSION

This paper presented and compared three real-time methods to psychoacoustically enhance the bass perception in audio signals using time, frequency and hybrid approaches for the generation of harmonics. The results of the subjective test were in line with the theory, highlighting that no algorithm performs better than the others in every scenario, but depending on bass tones, styles and patterns one specific approach is preferable. Moreover, better result can be achieved by simply tuning gain and cut-off parameters, to reach a desirable quality trade-off between bass enhancement and distortion.

5. REFERENCES

[1] E. Larsen and R. Aarts, “Audio bandwidth extension: Application of psychoacoustics, signal processing and loudspeaker design,” 2004.

[2] A. Hill and M. Hawksford, “A hybrid virtual bass system for optimized steady-state and transient performance,” 10 2010, pp. 1 – 6.

[3] J. Laroche and M. Dolson, “New phase-vocoder techniques for real-time pitch shifting,” 04 2000.

[4] M. Puckette, “Phase-locked vocoder,” in *Proceedings of 1995 Workshop on Applications of Signal Processing to Audio and Accoustics*, 1995, pp. 222–225.

[5] H. Mu, “Perceptual quality improvement and assessment for virtual bass systemr,” 2015, pp. 42–49.

[6] E. Moliner, J. Rämö, and V. Välimäki, “Virtual bass system with fuzzy separation of tones and transients,” 2020.

[7] J. Glover, V. Lazzarini, and J. Timoney, “Real-time detection of musical onsets with linear prediction and sinusoidal modeling,” 10 2011, pp. 1 – 6.

[8] “Recommendation itu-r bs.1534-1 - method for the subjective assessment of intermediate quality level of coding systems,” 2003.