VIRTUAL BASS SYSTEM WITH FUZZY SEPARATION OF TONES AND TRANSIENTS

Eloi Moliner, Jussi Rämö, and Vesa Välimäki *

Acoustics Lab, Dept. of Signal Processing and Acoustics
Aalto University
Espoo, Finland
eloi.moliner@aalto.fi

ABSTRACT

A virtual bass system creates an impression of bass perception in sound systems with weak low-frequency reproduction, which is typical of small loudspeakers. Virtual bass systems extend the bandwidth of the low-frequency audio content using either a nonlinear function or a phase vocoder, and add the processed signal to the reproduced sound. Hybrid systems separate transients and steady-state sounds, which are processed separately. It is still challenging to reach a good sound quality using a virtual bass system. This paper proposes a novel method, which separates the tonal, transient, and noisy parts of the audio signal in a fuzzy way, and then processes only the transients and tones. Those upper harmonics, which can be detected above the cutoff frequency, as boosted using timbre-matched weights, but missing upper harmonics are generated to assist the missing fundamental phenomenon. Listening test results show that the proposed algorithm outperforms previous methods in terms of perceived bass sound quality. The proposed method can enhance the bass sound perception of small loudspeakers, such as those used in laptop computers and mobile devices.

1. INTRODUCTION

Small loudspeakers have a limited frequency range in which they can operate efficiently. Low-frequency sound reproduction requires a large volume velocity, a condition that cannot be always satisfied due to the limitations of the loudspeaker. The lack of low-frequency components when reproducing music usually leads to a weak perception of bass and rhythm, as also drum sounds get suppressed. This paper studies methods to enhance the bass sounds, when a loudspeaker cannot reproduce low frequencies.

A conventional method to improve bass sounds is to boost the low-frequency range of the audio signal using an equalizing filter. However, a very high level of energy would be needed to boost the lowest frequencies, which usually leads to distortion and might cause unrecoverable damage to the loudspeaker. In practice, it is therefore often impossible to much improve the bass response of a small loudspeaker using a conventional equalizer.

A virtual bass system (VBS) enhances the bass perception by tricking the human auditory system to perceive low tones from higher harmonics. The goal of the VBS is to extend the low-frequency bandwidth so that the missing fundamental phenomenon [1, 2] will help people to perceive the low frequencies that are physically absent or significantly attenuated.

Early VBS systems distorted the low frequencies in the time domain. In the late 1990s, several patents suggested full and half-wave rectifiers for harmonic generation [3, 4, 5]. MaxxBass is one of the first commercial VBS products, and is based on nonlinear processing using a feedback multiplier [6, 7]. Gan *et al.* proposed a VBS using an amplitude modulator, which performed slightly better than MaxxBass [8]. Larsen and Aarts published a generic VBS framework using nonlinear devices (NLDs) [9] and also the first textbook on audio bandwidth extension [10], which analyzed the use of NLDs from a psychoacoustical viewpoint. Oo and Gan conducted an exhaustive study on the harmonic response and the intermodulation distortion of NLDs [11] and introduced novel NLDs, such as polynomial-based harmonic shifting [12].

Alternatively, bandwidth extension can be accomplished using frequency-domain processing. Bai *et al.* proposed the first VBS based on a phase vocoder (PV) [13]. They generated the harmonic series by pitch shifting the entire lowpass filtered signal. This method allows precise control over the weighting of the harmonic components, working better than most of NLDs for steady state signals. However, because of the temporal smearing effect caused by frame-based processing, this method performs poorly in transient-like sounds.

The successive strategy consisted in combining the NLD and the PV methods by separating the audio signal in transient and steady state components. In 2010, Hill [14] presented the first hybrid system having a transient detector based on the constant-Q transform to switch between the NLDs and the PV. Mu and Gan [15, 16] developed an improved hybrid VBS using the median filtering technique for transient and steady state audio separation [17].

The harmonics generated with a PV should be weighted in a way that not only enhances the bass but also generates a pleasant sound for the listener. Bai *et al.* [13] used a weighting based on the equal-loudness contours. Mu *et al.* presented a timbre matching approach [18], which consisted of weighting the harmonics with the spectral envelope of the original signal. Moon *et al.* proposed a phase-matched exponential weighting scheme [19], preserving the phases of the harmonics already existing in the original signal.

This paper proposes a bass enhancement method based on a fuzzy separation [20] of the transient, tonal, and noisy components of the audio signal, and on an appropriate processing of these components separately. In comparison to previous hybrid aproaches [14, 15, 16], we use the information extracted from the median filters to separate also the noise, leading to a better discrimination of transients and harmonic tones. The proposed method consists in an NLD for the transients and a refined phase vocoder processing for the tonal components, but no processing of the noisy part.

* This research is part of the Nordic Sound and Music Computing Network (NordicSMC), NordForsk project no. 86892.

Regarding the phase vocoder processing, we propose an improved harmonic enhancement technique based on the detection and preservation of the harmonics originally present in the signal. We uniquely generate by pitch-shifting the non-detected harmonics, while the detected ones only have their magnitudes scaled. We include a limited timbre-matching weighting scheme as well as an improved harmonic weighting methodology.

The rest of the paper is organized in the following way. Sec. 2 describes the idea of fuzzy separation of tonal, transient, and noise components in an audio signal. Sec. 3 discusses the transient processing and Sec. 4 the tonal processing in the proposed VBS. Sec. 5 presents a subjective listening test in which the proposed VBS is compared with previous methods. Sec. 6 concludes this paper.

2. FUZZY CLASSIFICATION OF SPECTRAL BINS

The separation method used in the proposed system is based on the median filter technique by Fitzgerald [17]. This method is based on the fact that, in a time-frequency representation, the tonal components of the signal tend to appear like ridges in the direction of the time axis while transient components appear as ridges in the direction of the frequency axis. In order to reduce the computational cost of the overall system, the lowpass filtered signal to be processed is first downsampled down to the sample rate of $f_\mathrm{s} = 4096\,\mathrm{Hz}$. Subsequently, the short-time Fourier Transform (STFT) of the downsampled signal is computed, with a frame size of 512 samples, no zero padding, a hop size of 64 samples and using a Hann window.

The generated time-frequency signal $X(m, k)$ is processed by the median filter technique in two ways:

$$X_\mathrm{s}(m, k) = \mathrm{med}\left(\left|X\left(m_0, k\right)\right|, ..., \left|X\left(m_L, k\right)\right|\right) \tag{1}$$

and

$$X_\mathrm{t}(m, k) = \mathrm{med}\left(\left|X\left(m, k_0\right)\right|, ..., \left|X\left(m, k_L\right)\right|\right), \tag{2}$$

where $X_\mathrm{s}(m, k)$ and $X_\mathrm{t}(m, k)$, respectively, are the tonal and transient STFTs, med() is the median filter operation, and m_0 and m_L are the starting and ending index in the time direction:

$$m_0 = m - \frac{L_\mathrm{t}}{2} + 1 \tag{3}$$

and

$$m_L = m + \frac{L_\mathrm{t}}{2}, \tag{4}$$

and L_t is the length of the median filter along the time axis. Parameters k_0 and k_L are the indices in the frequency direction:

$$k_0 = m - \frac{L_\mathrm{f}}{2} + 1 \tag{5}$$

and

$$k_L = m + \frac{L_\mathrm{f}}{2}, \tag{6}$$

where L_f is the length of the median filters along the frequency axis.

Consequently, the tonalness $R_\mathrm{s}(m, k)$ and transientness $R_\mathrm{t}(m, k)$ parameters can be defined in a similar way as in [20]. The tonalness and transientness, respectively, are defined as

$$R_\mathrm{s}(m, k) = \frac{|X_\mathrm{s}(m, k)|^2}{|X_\mathrm{s}(m, k)|^2 + |X_\mathrm{t}(m, k)|^2} \tag{7}$$

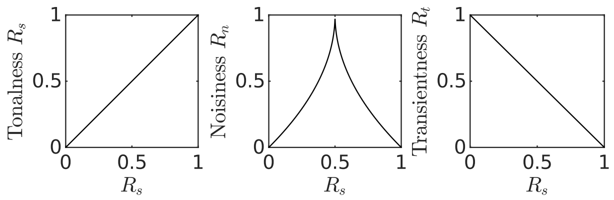

Figure 1: *Relations between the tonalness, noisiness, and transientness parameters used in this study.*

and

$$R_\mathrm{t}(m, k) = 1 - R_\mathrm{s}(m, k) = \frac{|X_\mathrm{t}(m, k)|^2}{|X_\mathrm{s}(m, k)|^2 + |X_\mathrm{t}(m, k)|^2}. \tag{8}$$

The noisy component is defined in [20] as those frequency bins where R_s and R_s tend to 0.5, in other words, where there is more uncertainty in the discrimination. However, here the noisiness parameter $R_\mathrm{n}(m, k)$ is defined in a different way as

$$R_\mathrm{n}(m, k) = 1 - \sqrt{|R_\mathrm{s}(m - k) - R_\mathrm{t}(m, k)|}. \tag{9}$$

The square root is used to reduce the energy of the R_n signal, such that only the spectral bins with R_s and R_t closer to 0.5 have higher values of noisiness. The relations between the three parameters are shown in Fig. 1.

These parameters are used as soft masks that will be applied to generate the transient $T(m, k)$, tonal $S(m, k)$, and noise $N(m, k)$ signals, in a similar way as in [15], but extending it to include the $N(m, k)$ signal:

$$S(m, k) = X(m, k)\left(R_\mathrm{s}(m, k) - \frac{1}{2}R_\mathrm{n}(m, k)\right), \tag{10}$$

$$T(m, k) = X(m, k)\left(R_\mathrm{t}(m, k) - \frac{1}{2}R_\mathrm{n}(m, k)\right), \tag{11}$$

and

$$N(m, k) = X(m, k)R_\mathrm{n}(m, k). \tag{12}$$

It should be noticed that the substraction of R_n from R_s and R_t in (10) and (11) is meant to ensure the perfect reconstruction of the signal, and thus

$$S(m, k) + T(m, k) + N(m, k) = X(m, k). \tag{13}$$

Figure 2 shows an example of the fuzzy separation result.

3. TRANSIENT PROCESSING WITH NLD

Figure 3 shows the block diagram of the proposed VBS. The fuzzy separated transient components are processed in the time domain with the generic framework based on NLDs from [9]. The transient signal $T(m, k)$ is transformed back into the time domain using the inverse short-time Fourier Transform (ISTFT).

The signal is thereafter split by lowpass and highpass filters LPF_t and HPF_t, respectively. They are FIR filters of order 500 and have the same cutoff frequency f_c, which is defined as the lowest frequency that the target loudspeaker is able to reproduce adequately. The filters were designed using the window method with the Chebyshev window. It is assumed that any frequency component below f_c will be highly attenuated by the loudspeaker and,

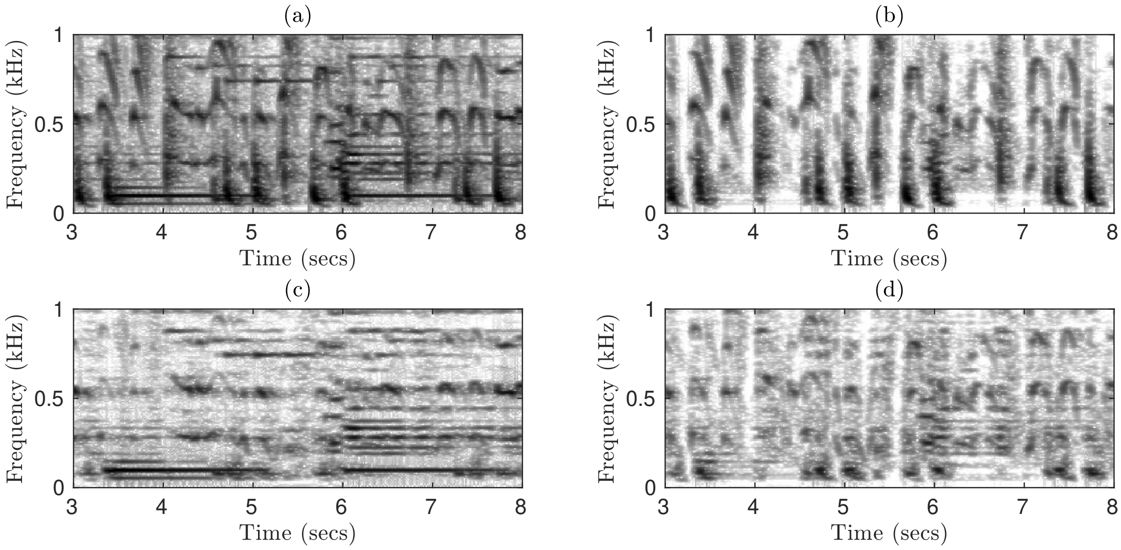

Figure 2: *Example of fuzzy separation of (a) input signal $X(m,k)$ into its (b) transient part $T(m,k)$, (c) tonal part $S(m,k)$, and (d) noisy part $N(m,k)$.*

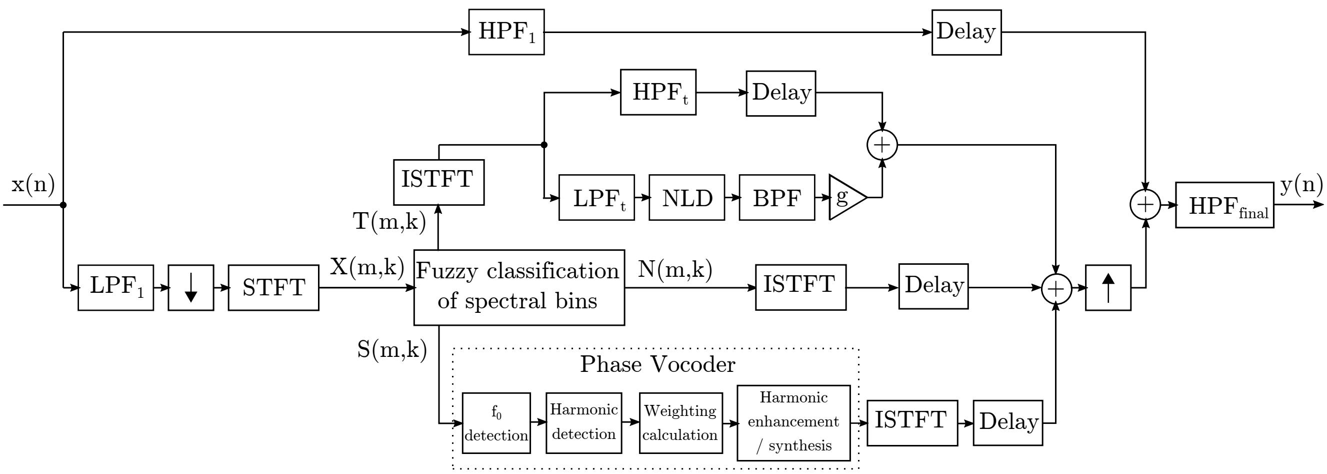

Figure 3: *Proposed virtual bass algorithm in which the transient part $T(m,k)$ is processed in a different way than the tonal part $S(m,k)$.*

therefore, it should be suppressed after performing the bandwidth extension. The f_c parameter depends on the frequency response of the loudspeaker, and its value may vary between about 100 Hz and 200 Hz in small loudspeakers.

Choosing an optimal nonlinear function for the VBS is not an obvious task, as was shown by Oo *et al.* who analyzed and graded different NLDs [21, 22]. The nonlinear function we apply is a half-wave rectifier cascaded with a fuzz exponential (HWR-FEXP1), having the transfer function shown in Fig. 4. The main motivation for using this NLD lays on the fact that the magnitude gaps between generated harmonics do not vary with the input magnitude [21]. Its use is not recommended in [22], because it shows poor performance in intermodulation distortion tests. However, since we are applying this NLD to transient signals, which are non-tonal, intermodulation distortion can be neglected.

The half-wave rectifier generates even harmonics due to its asymmetry. Because of this fact, the missing fundamental effect produces the perception of a fundamental at $2f_0$, one octave higher than desired. Therefore, we also apply the FEXP1, which only generates odd harmonics, to produce a complete harmonic series. This combination has proven to be effective for transient signals in other studies [16].

A bandpass filter (BPF) is applied after the NLD in Fig. 3 to remove all the low-frequency components below f_c and to attenuate the generated higher frequency harmonics to control the sharpness of the processed sound. Thus, the BPF has its lower cutoff frequency at f_c and its higher cutoff at $4f_c$. After that, a gain g is applied to amplify the bandpassed signal so it can have a perceivable loudness in reference to the lost low-frequency components. The g parameter can be adjusted with typical values between 8 dB and 12 dB, depending on the desired amount of enhancement. At the end, the processed signal is added to the highpassed signal,

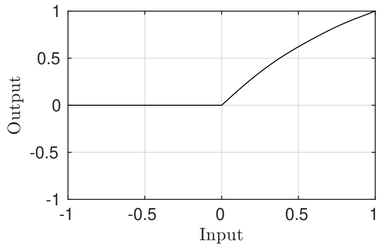

Figure 4: *HWR-FEXP1 transfer function is a combination of a half-wave rectifier and an exponential function.*

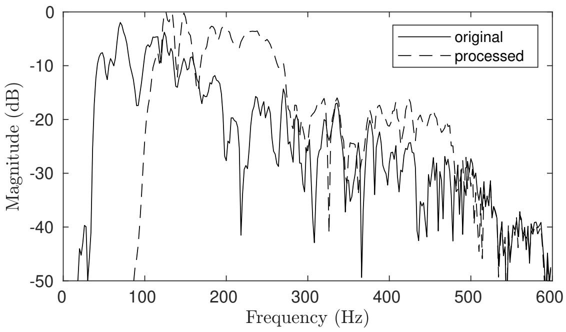

Figure 5: *Effect of NLD processing on the spectrum of a kick drum sound, when the cutoff frequency is 120 Hz.*

taking the delay difference caused by the BPF into account.

Figure 5 shows the input and output spectra of a kick drum sample. It is seen that the processed signal is lacking the frequencies below the cutoff frequency 120 Hz. However, the frequencies around 200 Hz have been amplified by the NLD processing.

4. TONAL PROCESSING WITH PHASE VOCODER

Tonal components are processed in the frequency domain by using a phase vocoder, as shown in the lowest branch in Fig. 3. The PV processing can be divided into three main parts: the detection of the f_0 component and its respective harmonics, the calculation of the harmonic weighting and, finally, the harmonic enhancement.

4.1. Fundamental and Harmonic Detection

The first step is a relaxed spectral peak detection. All local maxima in the magnitude spectrum are selected as peaks, the only constraints to be made in the search are a frequency range and a magnitude threshold. The frequency range is selected between $\frac{1}{4}f_c$ and $\frac{1}{2}f_s$. The threshold parameter can be tuned depending on the dynamic range of the audio signal. For a more precise peak location, its position and magnitude values are estimated using parabolic interpolation [23].

The maximum peak below f_c from the set of detected peaks is selected as the fundamental candidate. Afterwards, it is studied whether the candidate is the true f_0 component or whether it could be the second harmonic of another detected peak in a lower frequency. In the latter case, the lower peak is selected as the definitive f_0.

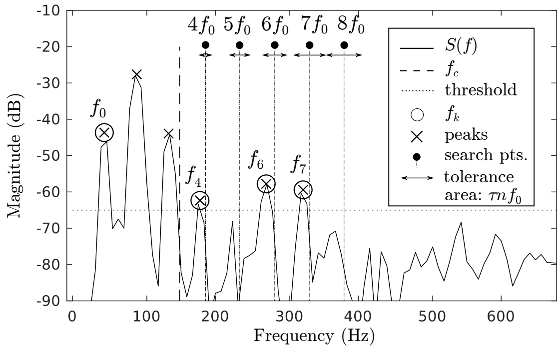

Figure 6: *Example of harmonic detection when* $f_0 \in [\frac{1}{4}f_c, \frac{1}{3}f_c]$.

Once the f_0 component has been detected, the following step is to search all its respective harmonics, considering each harmonic to be a multiple of f_0:

$$f_k = kf_0, \tag{14}$$

where $k = 1, 2, 3...$ is the order of each harmonic. Fixing the number of processed harmonics as K and adding the restriction that all of them should be located above f_c, it can be seen that the order of the harmonics to process depends on which interval f_0 is located:

$$f_k = \begin{cases} f_2, f_3, ..., f_{K+1}, & \text{if } f_0 \in [\frac{1}{2}f_c, f_c], \\ f_3, f_4, ..., f_{K+2}, & \text{if } f_0 \in [\frac{1}{3}f_c, \frac{1}{2}f_c], \\ f_4, f_5, ..., f_{K+3}, & \text{if } f_0 \in [\frac{1}{4}f_c, \frac{1}{3}f_c]. \end{cases} \tag{15}$$

Each harmonic is searched by the statement

$$f_k = \arg\min_f(|(\forall f \in \text{peak locs}) - kf_0|) \tag{16}$$

and is detected as a harmonic, if the following condition is satisfied:

$$|f_k - kf_0| < \tau k f_0, \tag{17}$$

where τ is the tolerance parameter, meant to relax the search for the cases in which the signal is not precisely harmonic, like in string instrument sounds [24]. Small values of τ will search and posteriorily generate more accurately harmonic partials, while large values will tend to find more inharmonic partials and preserve a natural spectral structure. The drawback of using a large τ is that the probability of detecting a wrong harmonic increases, e.g. a peak belonging to another tone overlapping in time with the bass tone. A common value used in this work is $\tau = 0.05$.

4.2. Limited Timbre Matching Weighting Scheme

We introduce a novel weighting scheme based on the timbre matching approach of Mu *et al.* [18]. This method tends to weight the generated harmonics in a way that the spectral envelope of the original signal is preserved and, consequently, the perceived timbre is similar.

We compute an estimation of the spectral envelope on each frame by applying a Bark scale triangular filter bank to the signal. Each band is calculated as

$$\text{Band}_j = \frac{\sum_{k\in j} \text{TF}_j(k)|S(k)|}{\sum_{k\in j} \text{TF}_j(k)}, \tag{18}$$

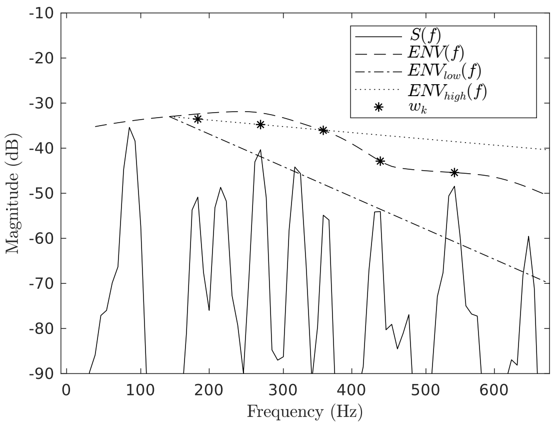

Figure 7: *Example of weighting of harmonics. The cutoff frequency of the highpass filter is 150 Hz.*

where TF_j represents the triangular filter centered at the band j and $S(k)$ is the tonal spectrogram. To generate the complete envelope $\mathrm{ENV}(f)$, the resulting band magnitudes are interpolated with cubic interpolation. The resulting envelope is scaled up to the magnitude of the f_0 component:

$$\mathrm{ENV}(f) = \mathrm{ENV}(f)\frac{|S(f_0)|}{\mathrm{ENV}(f_0)}, \tag{19}$$

where $\mathrm{ENV}(f_0)$ is the magnitude of the spectral envelope at f_0 and $S(f_0)$ is the tonal spectrogram at f_0. In the case that a first or second harmonic below f_c has a bigger magnitude than f_0, the envelope is be scaled up to the magnitude of this harmonic.

Mu *et al.* [18] used the interpolated result to weight the generated harmonics, but this approach leads to two problems. Sometimes the estimated envelope has a very strong decay fitted to the original signal and, using this weighting, almost no enhancement is obtained. In an opposite case, the envelope can be altered by an interference signal, such as another tone of a higher frequency overlapping in time with the bass, and higher volume harmonics can be unnecessarily generated creating undesired effects.

To mitigate these effects, the envelope is limited between two constant exponentially decaying curves:

$$\mathrm{ENV}_{\mathrm{low}}(f) = \mathrm{ENV}(f_\mathrm{c}) - \alpha_{\mathrm{low}}\frac{f}{f_\mathrm{c}} \tag{20}$$

and

$$\mathrm{ENV}_{\mathrm{high}}(f) = \mathrm{ENV}(f_\mathrm{c}) - \alpha_{\mathrm{high}}\frac{f}{f_\mathrm{c}}, \tag{21}$$

where $\mathrm{ENV}_{\mathrm{low}}(f)$ and $\mathrm{ENV}_{\mathrm{high}}(f)$ are, respectively, the lower and upper limits in dB. Parameters α_{low} and α_{high} are the attenuation factors in dB/oct. Appropriate values of these factors are $\alpha_{\mathrm{low}} = 10$ dB/oct and $\alpha_{\mathrm{high}} = 3$ dB/oct. Therefore, each harmonic f_k is weighted by:

$$w_k = \begin{cases} \mathrm{ENV}_{\mathrm{low}}(f_k), & \text{if } \mathrm{ENV}_{\mathrm{low}}(f_k) > \mathrm{ENV}(f_k), \\ \mathrm{ENV}_{\mathrm{high}}(f_k), & \text{if } \mathrm{ENV}_{\mathrm{high}}(f_k) < \mathrm{ENV}(f_k), \\ \mathrm{ENV}(f_k) & \text{otherwise}, \end{cases} \tag{22}$$

where w_k is the weight corresponding to each harmonic f_k. This way we can assure a minimum weighting, defined by $\mathrm{ENV}_{\mathrm{low}}$, and a maximum, defined by $\mathrm{ENV}_{\mathrm{high}}$. This weighting methodology is illustrated in Fig. 7.

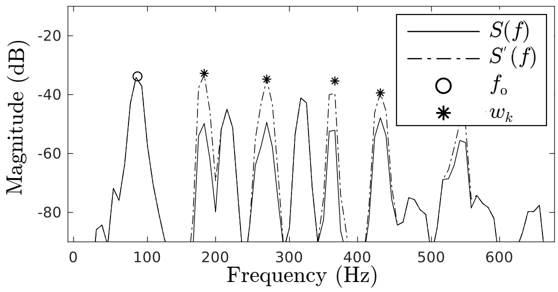

Figure 8: *Enhancement of detected harmonics.*

4.3. Harmonic Enhancement

Most VBS aproaches generate the partials by simply pitch shifting the low-frequency components without preserving the original phase spectrum [13, 14, 15]. However, several studies have indicated that changes in the phase spectrum affect the perception of timbre [25, 26, 27]. Furthermore, phase non-coherence is common in acoustic sounds and preserving it would contribute to the naturalness of the audio perception [28]. In order to minimally alter the timbre of the signal, we follow a new strategy consisting in preserving the original phase spectrum as well as possible.

The harmonics that have been detected in Sec. 4.1 are not resynthesized but they are enhanced. Their magnitude is scaled according to its targeted weight w_k from Sec. 4.2 while the phase of these harmonics is not modified. To perform the magnitude scaling, we apply a frequency-domain window ω_{ROI} to isolate the region of influence (ROI) of the peak. A Tukey window with its size slightly larger than the main lobe width of the analysis window is chosen, so it does not alter the magnitude on the top of the lobe but guarantees a smooth addition on the sides. The detected harmonics H_k are enhanced by:

$$H_k = \frac{w_k}{|S(f_k)|}S(f)\omega_{\mathrm{ROI}}(f - f_k). \tag{23}$$

Figure 8 shows an example of magnitude enhancement of the detected harmonics.

This methodology is similar to Moon's phase-matched exponential harmonic weighting [19]. However, Moon's method consisted in pitch shifting the magnitudes of the harmonics without changing the phases, not considering whether there was already a partial in that location or not.

The rest of the partials that have not been previously detected are generated by pitch shifting using the technique from [29]. The same window ω_{ROI} is applied to the f_0 and its neighboring bins. Each harmonic frequency is simply calculated as a multiple of f_0, as defined in (14), and the ROI of the fundamental is pitch shifted to the modified spectrum at this exact frequency with its corresponding weight w_k. Considering the rounding errors on shifting FFT bins, in order to have a more precise shifting to the exact target frequency, a fractional delay filter based on a Lagrange interpolator is applied.

The phases of each shifted partial should be recomputed in order to maintain the phase-coherence between frames. This can

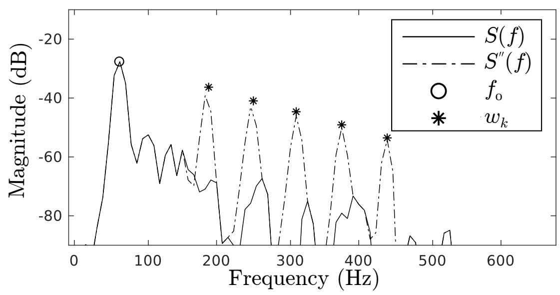

Figure 9: *Harmonic generation by pitch shifting, when harmonics of a fundamental have not been detected.*

be performed by multiplying the frequency bins of the shifted ROI by the complex factor $Z_{u,k}$:

$$Z_{u,k} = Z_{u-1,k} e^{j2\pi(f_k - f_0)R}, \tag{24}$$

where R is the hop size and Z_{u-1} is the value from the previous iteration. Each non-detected harmonic H_k is generated by pitch shifting following the equation:

$$H_k = \left[\frac{w_k Z_{u,k}}{|S(f_0)|} S(f) \omega_{\text{ROI}}(f - f_0) \right] * \delta(f - kf_0), \tag{25}$$

where $\delta(f)$ is the Dirac delta function and $*$ is the convolution operator. Figure 9 shows an example where all the harmonics are generated by pitch shifting.

Finally, the modified spectra $S'(f)$ can be constructed by adding both the enhanced and the pitch-shifted harmonics to the spectrum:

$$S'(f) = S(f) + \sum_{k}^{K} H_k - S(f)\omega_{ROI}(f - f_k). \tag{26}$$

5. SUBJECTIVE LISTENING TEST

The MUSHRA (MUltiple Stimuli with Hidden Reference and Anchor) [30] test was chosen for the blind comparison of different VBS methods. The audio material selected for the listening test and listed in Table 1 contains variable bass content. The rock song has electric bass tones whereas the jazz song contains an acoustic bass but no other low-frequency sounds. The hip hop song features long tones of electric bass and a very loud and boomy bass drum. The classical music example contains a deep bass tone. All these audio excerpts are dramatically affected by highpass filtering.

The audio samples were processed using four selected methods, including the hybrid method proposed in this paper, an NLD-based VBS [22], a PV-based VBS [13], and Hill's hybrid VBS [14]. All of them were highpass filtered with the cutoff frequency of 150 Hz. The parameters of all four methods were tuned to obtain the best performance for each stimulus. Regarding our method, we adjusted the number of processed harmonics, the weighting limits α_{low} and α_{high}, the gain g for the transient processing and the magnitude threshold for the harmonic detection, which depends on the dynamic range of the signal. The harmonic gain parameter was adjusted for the NLD condition, as well as for the PV, where the number of shifted harmonics was also tuned. Hill's hybrid method samples were generated using the toolbox published by himself in his website [31], which let us modify the number of processed harmonics and both transient and tonal gains.

Table 1: *Audio examples used in the listening test.*

	Genre	Artist	Title	Time
1	Hip hop	Wu-Tang Clan	C.R.E.A.M	0:22
2	Jazz	Miles Davis	So What	0:52
3	Rock	Radiohead	Karma Police	1:43
4	Classical	Richard Strauss (Berlin Philharmonic)	Also Sprach Zarathustra	1:03

The original, unprocessed signal was used as a reference in each case, and was also included among the MUSHRA test items as a hidden reference. The anchor, or a low-quality signal specifying the low end of the perceptual scale, was a highpass filtered version of the original signal with a 150-Hz cutoff frequency.

The subjective test was performed using webMUSHRA [32], a web-based interface compliant with the ITU-R Recommendation BS.1534 [30]. The listeners were asked to complete the test using headphones without restrictions. The listeners were asked to evaluate the quality of the bass perception on a scale from 0 to 100, in comparison to the reference sound, for all six items, which appeared on the same page without labels in random order: anchor, hidden reference, and the four processed signals. All four test signals were included twice, so that the MUSHRA test contained eight pages, and the subjects had to evaluate 48 audio files in total. The duration of the test items was about 11 s, but the subjects could select to repeat a shorter segment of each file by adjusting the beginning and end points using sliders.

Altogether 18 people participated in the listening test. However, the data of 7 subjects were excluded, since they rated the hidden reference condition below 90 points in more than 15% of the test items, as this was recommended in [30]. From the 11 included listeners, three were females and nine had previous experience of formal listening tests, and their average age was 28 years. The test took usually about 20 min, and was not considered too tiresome by the subjects.

5.1. Results of Subjective Evaluation

Figure 10 shows the mean results of the MUSHRA test with 95% confidence interval for the different audio items. The mean scores and their overall averages are also shown in Table 2. All conditions except the reference received relatively low scores, often rated as fair or poor, due to their contrast with the reference, which was the only signal containing bass frequencies below 150 Hz. The listeners were able to easily distinguish the hidden reference, which received nearly 100 points most of the time, and, consequently, they gave consistently lower ratings to the other conditions. The anchor had the lowest rating in all genres expect in jazz.

The proposed hybrid method received the best mean score after the reference for all four audio samples, having a clear difference in jazz and rock, but a tighter result in the other genres. The PV received a mean score very close to our proposed method in all cases except jazz, where the temporal smearing of transients, caused by the PV processing, was more easily perceivable than in other signals. In some cases, the smearing was masked by other musical instruments, and the listeners did not recognize it as a dis-

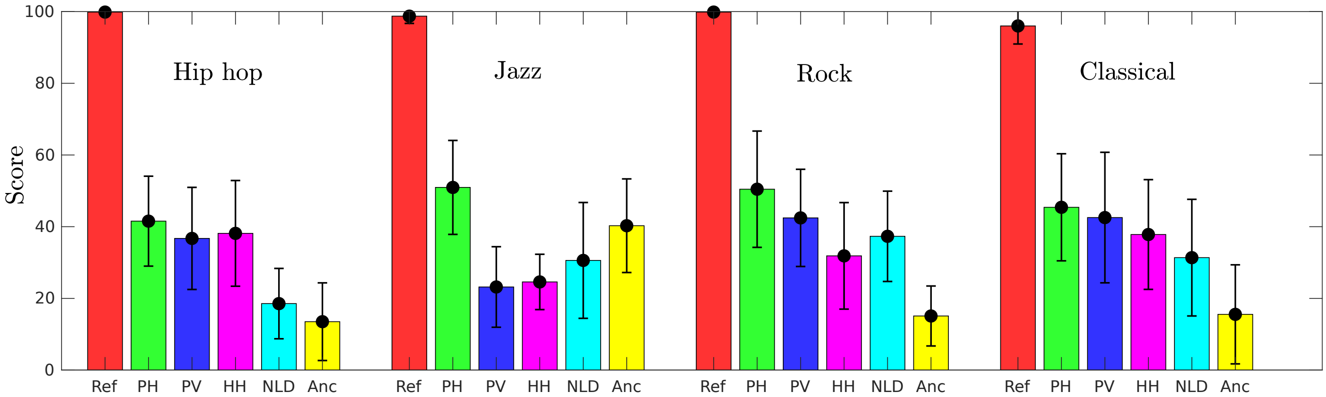

Figure 10: *Mean results of MUSHRA test with 95% confidence intervals for all the audio samples and conditions (Ref = reference, PH = proposed hybrid, NLD = Non-Linear Device, PV = Phase Vocoder, HH = Hill's Hybrid, Anc = anchor).*

Table 2: *Mean MUSHRA scores for test items reference (Ref), proposed hybrid (PH), Non-Linear Device (NLD), Phase Vocoder (PV), Hill's hybrid (HH), and anchor (Anc) for all genres. The best score (excluding Ref) on each line is highlighted.*

Genre	**Ref**	**PH**	**PV**	**HH**	**NLD**	**Anc**
Hip hop	99.8	**41.5**	36.7	38.1	18.5	13.5
Jazz	98.7	**50.9**	23.1	24.5	30.5	40.2
Rock	99.8	**50.4**	42.4	31.8	37.3	15.0
Classical	96.0	**45.4**	42.5	37.8	31.3	15.5
Average	98.6	**47.0**	36.2	33.1	29.4	21.1

turbing artifact. Hill's hybrid system did not achieve significantly better results than the low-quality anchor, showing that its transient separation technique does not perform as well as expected.

In some examples, such as hip hop and classical, the differences are not statistically significant as the confidence intervals are overlapping. Nevertheless, in average, the results show that the proposed method outperforms the rest of techniques we compared it with.

It can be speculated that the wide confidence intervals shown in Fig. 10 may have been caused by the variance in the headphone quality of the subjects. Listeners with high-fidelity headphones would be able to hear the lowest frequency range of the reference signal, while people with lower quality headphones might not hear them that well and give higher ratings to the other conditions. Alternatively, it is possible that the listeners had difficulties in forming a clear opinion about the different methods, as they were not familiar with audio distortion caused by various bass enhancement algorithms.

The audio items used in the test will be available online at: http://research.spa.aalto.fi/publications/papers/dafx20-vbs.

6. CONCLUSION

This paper presented a virtual bass system based on a novel transient, tonal, and noise component separation technique, incorporating also an improved processing methodology for the tonal components. Our motivation was to design a VBS that not only enhances the bass frequencies but also tends to preserve the original timbre of the signal with a minimal distortion. The proposed algorithm was compared with several other methods, the NLD, PV, and Hill's hybrid system, by performing a formal listening test. The test scores verified that our method produces a better perceived audio quality in the bass range than the compared methods. The new VBS can be applied to audio reproduction with small loudspeakers to enhance deep bass and drum sounds, which would otherwise be inaudible.

7. ACKNOWLEDGMENTS

This work was conducted during Eloi Moliner's visit to Aalto University, where he was working on his Master's thesis in January–July 2020 with the support of the ERASMUS program and the Universitat Politècnica de Catalunya, Barcelona, Spain. Many thanks to the listeners who took part in the subjective test. The authors also thank Adam J. Hill for publishing his Virtual Bass Toolbox.

8. REFERENCES

[1] J.F. Schouten, R.J. Ritsma, and B.L. Cardozo, "Pitch of the residue," *J. Acoust. Soc. Am.*, vol. 34, no. 9, pp. 1418, 1962.

[2] B.C.J. Moore, *An Introduction to the Psychology of Hearing*, Emerald, 2012.

[3] M. Oda, "Low frequency audio conversion circuit," US Patent 5,668,885, Sept. 16, 1997.

[4] T. Unemura, "Audio circuit," US Patent 5,771,296, June 23, 1998.

[5] E. E. Feremans and F. De Smet, "Method and device for processing signals," US Patent 5,828,755, Oct. 27 1998.

[6] D. Ben-Tzur, "The effect of the MaxxBass 1 psychoacoustic bass enhancement system on loudspeaker design," in *Proc. AES 106th Conv.*, 1999.

[7] M. Shashoua and D. Glotter, "Method and system for enhancing quality of sound signal," US Patent 5,930,373, July 27, 1999.

[8] W.-S. Gan, S. M. Kuo, and C. W. Toh, "Virtual bass for home entertainment, multimedia PC, game station and portable audio systems," *IEEE Trans. Consumer Electronics*, vol. 47, no. 4, pp. 787–796, Nov. 2001.

[9] E. Larsen and R. Aarts, "Reproducing low-pitched signals through small loudspeakers," *J. Audio Eng. Soc.*, vol. 50, no. 3, pp. 147–164, Jan. 2002.

[10] E. Larsen and R.M. Aarts, *Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design*, Wiley, 2005.

[11] N. Oo and W.-S. Gan, "Harmonic analysis of nonlinear devices for virtual bass system," in *Proc. Int. Conf. Audio, Language and Image Processing*, Aug. 2008, pp. 279–284.

[12] W.-S. Gan and N. Oo, "Harmonic and intermodulation analysis of nonlinear devices used in virtual bass systems," in *Proc. AES 124th Conv.*, May 2008.

[13] M. Bai and C. Lin, "Synthesis and implementation of virtual bass system with a phase-vocoder approach," *J. Audio Eng. Soc.*, vol. 54, 11 2006.

[14] A. J. Hill and M. O. J. Hawksford, "A hybrid virtual bass system for optimized steady-state and transient performance," in *Proc. Computer Science and Electronic Engineering Conf. (CEEC)*, Sep. 2010, pp. 1–6.

[15] H. Mu, W.-S. Gan, and E. Tan, "A psychoacoustic bass enhancement system with improved transient and steady-state performance," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2012, pp. 141–144.

[16] H. Mu and W.-S. Gan, "Perceptual quality improvement and assessment for virtual bass systems," *J. Audio Eng. Soc.*, vol. 63, no. 11, pp. 900–913, 2015.

[17] D. Fitzgerald, "Harmonic/percussive separation using median filtering," *Proc. Int. Conf. Digital Audio Effects (DAFx-10)*, Sept. 2010.

[18] H. Mu, W.-S. Gan, and E. Tan, "A timbre matching approach to enhance audio quality of psychoacoustic bass enhancement system," in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Process.*, May 2013, pp. 36–40.

[19] H. Moon, G. Park, Y. Park, and D. H. Youn, "A phase-matched exponential harmonic weighting for improved sensation of virtual bass," in *Proc. AES Conv. 140*, May 2016.

[20] E.-P. Damskägg and V. Välimäki, "Audio time stretching using fuzzy classification of spectral bins," *Applied Sciences*, vol. 7, pp. 1293, Dec. 2017.

[21] W.-S. Gan and N. Oo, "Analytical and perceptual evaluation of nonlinear devices for virtual bass system," in *Proc. AES 128th Conv.*, May 2010.

[22] N. Oo, W.-S. Gan, and M.O.J. Hawksford, "Perceptually-motivated objective grading of nonlinear processing in virtual bass systems," *J. Audio Eng. Soc.*, vol. 59, no. 11, pp. 804–824, Nov. 2011.

[23] J. O. Smith and X. Serra, "PARSHL: An analysis/synthesis program for non-harmonic sounds based on a sinusoidal representation," in *Proc. Int. Computer Music Conf.*, Urbana, IL, USA, 1987, pp. 290–297.

[24] H. Järveläinen, V. Välimäki, and M. Karjalainen, "Audibility of the timbral effects of inharmonicity in stringed instrument tones," *Acoustics Research Letters Online*, vol. 2, no. 3, pp. 79–84, Jul. 2001.

[25] R. Plomp and H. J. M. Steeneken, "Effect of phase on the timbre of complex tones," *J. Acoust. Soc. Am.*, vol. 46, no. 2B, pp. 409–421, 1969.

[26] B. C. J. Moore, "Interference effects and phase sensitivity in hearing," *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, vol. 360, pp. 833–858, 2002.

[27] M. Laitinen, S. Disch, and V. Pulkki, "Sensitivity of human hearing to changes in phase spectrum," *J. Audio Eng. Soc.*, vol. 61, no. 11, pp. 860–877, Nov. 2013.

[28] S. Dubnov and X. Rodet, "Investigation of phase coupling phenomena in sustained portion of musical instruments sound," *J. Acoust. Soc. Am.*, vol. 113, no. 1, pp. 348–359, 2003.

[29] J. Laroche and M. Dolson, "New phase-vocoder techniques for pitch-shifting, harmonizing and other exotic effects," in *Proc. IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, 1999, pp. 91–94.

[30] ITU-R, "Method for the subjective assessment of intermediate quality level of audio systems," Recommendation BS.1534-3, International Telecommunication Union, Geneva, Switzerland, Oct. 2015.

[31] "Vb toolbox," `http://adamjhill.com/vb-toolbox/`, Accessed: 2020-04-18.

[32] M. Schoeffler, S. Bartoschek, F. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webMUSHRA — A comprehensive framework for web-based listening tests," *J. Open Research Software*, vol. 6, Feb. 2018.