Virtual Bass Enhancement Via Music Demixing

Riccardo Giampiccolo, *Graduate Student Member, IEEE*, Alessandro Ilic Mezza, *Graduate Student Member, IEEE*, Alberto Bernardini, *Member, IEEE*, and Augusto Sarti, *Senior Member, IEEE*

***Abstract*—Virtual Bass Enhancement (VBE) refers to a class of digital signal processing algorithms that aim at enhancing the perception of low frequencies in audio applications. Such algorithms typically exploit well-known psychoacoustic effects and are particularly valuable for improving the performance of small-size transducers often found in consumer electronics. Though both time- and frequency-domain techniques have been proposed in the literature, none of them capitalizes on the latest achievements of deep learning as far as music processing is concerned. In this letter, we propose a novel time-domain VBE algorithm that incorporates a deep neural network for music demixing as part of the processing pipeline. This technique is shown to improve the bass perception and reduce inharmonic distortion, i.e., the main issue of existing time-domain VBE algorithms. The results of a perceptual test are then presented, showing that the proposed method is able to outperform state-of-the-art algorithms both in terms of bass enhancement and basic audio quality.**

***Index Terms*—Virtual bass enhancement, music demixing, psychoacoustics, perceptual test.**

I. Introduction

SINCE the beginning of digital audio signal processing, psychoacoustics has been instrumental for designing algorithms and protocols compliant with the human perception of sound [1]. Suffice to think of lossy compression formats, such as MP3, and their pervasive presence in digital audio applications [2]. More recently, research interest has also been geared towards psychoacoustic effects for enhancing the acoustical performance of small-size loudspeakers [3]. In fact, due to their reduced thickness, small-size loudspeakers are unable to provide high volume velocities, impairing especially the reproduction of low frequencies. In particular, the "missing fundamental" phenomenon is exploited to trick the human brain into perceiving low tones. According to such a phenomenon, a pitch related to a fundamental frequency f_0 is perceived not only if it is part of the audio track but also thanks to the periodicity of its higher harmonics ($2f_0$, $3f_0$, $4f_0$, etc.) [3].

In the literature, the algorithms that exploit such an effect are called *Virtual Bass Enhancement* (VBE) algorithms [3]–[6] and, depending on the technique involved in the harmonic generation, are typically categorized into time-domain, frequency-domain, and hybrid methods. On the one hand, time-domain techniques can be thought of as composed of three building blocks [5], [7], [8]: (i) a crossover network for the separation of low- and high-frequency components; (ii) a nonlinear device (NLD) for the generation of overtones, which acts only on the lowpass-filtered track; (iii) an output stage, which sums back the harmonically-enriched and the highpass-filtered track. On the other hand, frequency-domain techniques are typically based on phase vocoders, and employ pitch shifting for mapping low frequencies to regions of the spectrum wherein the loudspeaker is able to operate with full capability [9], [10]. Time-domain methods usually perform better on transients than on the tonal parts of audio tracks, whereas the opposite holds for frequency-domain methods. Hybrid techniques have been introduced with the purpose of merging the benefits of the two approaches [6], [11]. However, the latter are characterized by a high computational cost that prevents them from running in real-time, and thus, they are usually not tabled for consumer electronics applications.

The aforementioned methods make use of common digital signal processing techniques. Recently, in [12], the authors proposed to apply statistical machine learning methods for music genre recognition to select the best NLD for processing a certain audio track. However, to the best of our knowledge, no publication involves machine learning/deep learning techniques directly in the processing chain of VBE algorithms.

In this letter, we propose to exploit the latest advancements of deep learning as far as music processing is concerned to overcome some limitations of common approaches. For instance, time-domain techniques, although efficient and lightweight, suffer from the generation of inharmonic distortion, mainly due to the nonlinear processing of the superposition of multiple sound sources present in polyphonic tracks. Hence, we propose to substitute the crossover network, which is typically present at the front-end of time-domain VBE algorithms, with a deep-learning-based Music Demixing Model (MDM) [13]–[15] such that the NLD can be applied to a monophonic track (e.g., a single instrument) at a time. We then present a general algorithm, designed to be independent of the MDM of choice, and we introduce a new NLD as well as a new processing pipeline. Finally, a perceptual test proves the proposed method able to outperform state-of-the-art techniques in terms of both bass enhancement and Basic Audio Quality (BAQ).

II. Proposed Method

Fig. 1 shows the general block diagram of the proposed VBE algorithm. In the first place, the target audio frame $\mathbf{x} = [x[k], \dots, x[k+K-1]]^{\mathrm{T}}$, where K is the number of frame samples and k is the sample index, is processed by an MDM which extracts S stems, each one related to a different instrument or mixture of instruments that share a common sound

This work was funded and supported by INVENTVM Semiconductor s.r.l., Via Alessandro Brambilla, 60, 27100 Pavia, Italy. *(R. Giampiccolo and A. I. Mezza contributed equally to this work.) (Corresponding authors: R. Giampiccolo; A. I. Mezza.)*

The authors are with the Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano, Piazza Leonardo Da Vinci, 32, 20133 Milan, Italy (e-mails: {riccardo.giampiccolo, alessandroilic.mezza, alberto.bernardini, augusto.sarti}@polimi.it).

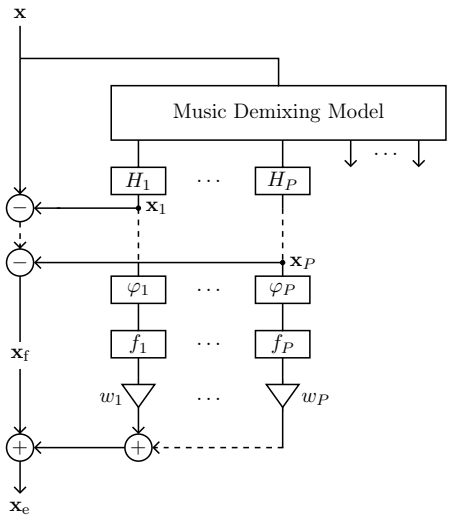


Fig. 1. Block diagram of the proposed VBE algorithm.

production mechanism. Then, P stems containing low frequencies are filtered by means of $H_1(z), \ldots, H_P(z)$; the remaining stems are, instead, discarded. The filtered stems $\mathbf{x}_1, \ldots, \mathbf{x}_P$ are then subtracted from the input fame $\mathbf{x}$ (obtaining $\mathbf{x}_\mathrm{f}$) and passed through two types of functions: normalization functions $\varphi_1, \ldots, \varphi_P$ and nonlinear devices $f_1, \ldots, f_P$. The processed stems are finally weighted by $w_1, \ldots, w_P$ and summed back to $\mathbf{x}_\mathrm{f}$, yielding the bass-enhanced audio frame $\mathbf{x}_\mathrm{e}$. The overall processing pipeline can be thus written as follows

$$\mathbf{x}_\mathrm{e} = \mathbf{x} + \sum_{p=1}^{P} \left(w_p \, \mathbf{f}_p(\varphi(\mathbf{x}_p)) - \mathbf{x}_p \right), \quad p = 1, \ldots, P, \quad (1)$$

where $\mathbf{x}_p \in \mathbb{R}^K$ is the vector containing the time-domain samples of the output of filter $H_p(z)$, whereas $\mathbf{f}_p(\varphi(\mathbf{x}_p)) = [f_p(\varphi(x_p[k])), \ldots, f_p(\varphi(x_p[k+K-1]))]^\mathrm{T}$.

In the next sections, we will analyze each part of the proposed algorithm, providing the details of our implementation.

A. Music Demixing Model and Filtering Stage

The algorithm is formalized to accommodate whichever MDM, existing or yet to be developed. In fact, although state-of-the-art MDMs are able to extract up to $S = 6$ stems (typically, `bass`, `drums`, `piano`, `vocals`, `guitar`, and `others`), in the future, more sophisticated models could be characterized by a higher S, providing additional stems that could come in handy for enhancing the bass perception.

In our implementation, we select Spleeter by Deezer as MDM [13]. Although it is no longer the state-of-the-art method as far as music source separation is concerned [14], [15], it is able to separate a mix audio track into 4 stems 100 times faster than real-time on a single Graphics Processing Unit (GPU): a computational speed yet to be matched by more recent MDMs. In particular, Spleeter consists of pre-trained 12-layer U-Nets [16] (one for each stem) employed to estimate spectro-temporal soft masks suitable to separate single sources [13]. Going more into detail, we set $S = 4$ and $P = 2$, processing thus only `bass` (marked with subscript 1) and `drums` (marked with subscript 2.) We then set $H_1(z) = 1$,

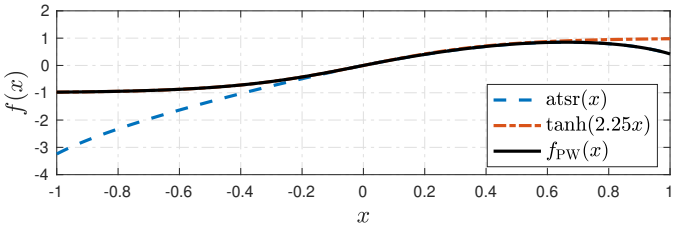


Fig. 2. Proposed piecewise nonlinear device $f_\mathrm{PW}(x)$ (in black) and its two sub-functions: $\mathrm{atsr}(x)$ (in blue) and $\tanh(2.25x)$ (in orange).

since we want to process the entire frequency range of the `bass` stem, and we employ a 10th order zero-phase forward-backward lowpass filter $H_2(z)$ [17] with cut-off frequency 250 Hz to extract the low end (kick drum and toms) from the `drums` stem. Zero-phase filters are considered to avoid phase distortion, which may impair the final result. In fact, as mentioned before, we aim at summing back each processed stem to the target audio frame $\mathbf{x}_\mathrm{f}$, and, therefore, we must guard against phase delays. It is worth noting that the same could have also been achieved by means of linear-phase filters, as long as proper delays are introduced in the pipeline.

Finally, the outputs of the filters are fed into normalization functions, which will be presented further ahead in Section II-C.

B. Proposed Nonlinear Device

The "missing fundamental" phenomenon occurs only when a minimum amount of harmonics is generated [3]. Moreover, the quality of bass enhancement strictly depends on the NLD selected for the harmonic generation [5]. Hence, over the past ten years, research effort has been put into deriving guidelines for selecting those NLDs characterized by good performance in terms of both bass enhancement and audio quality (low generation of inharmonic distortion) [5], [7]. In particular, [5] shows that the best NLDs are those characterized by neither odd nor even functions (and thus able to generate both odd and even harmonics), with a continuous first derivative, and a second derivative less than zero over the interval $(0, 1]$. For instance, functions like

$$\mathrm{atsr}(x) := 2.5 \tan^{-1}(0.9x) + 2.5\sqrt{1 - (0.9x)^2} - 2.5 \quad (2)$$

and $\tanh(x)$ are classified as "good" NLDs [5]. Nevertheless, $\tanh(x)$ is an odd function, and thus it is able to generate only odd harmonics, whereas $\mathrm{atsr}(x)$, although being neither even nor odd, can lead to highly asymmetric waveforms since it weighs negative inputs much heavier than positive inputs, as can be appreciated by looking at Fig. 2. In turn, this can hinder the amount of gain that can be safely applied to the target audio frame before clipping the negative half wave.

Given that audio signals take values in a bounded range $\mathcal{X}$ (typically $[-1, 1]$), we would like to have, instead, an NLD $f : \mathcal{X} \to \mathcal{Y}$ such that $\mathcal{Y} \subseteq \mathcal{X}$, in order to prevent loss of headroom and additional distortions.

Hence, we propose a novel nonlinear device f_PW defined as

$$f_\mathrm{PW}(x) := \begin{cases} \mathrm{atsr}(x), & \text{if } x \geq 0 \\ \tanh(2.25x), & \text{if } x < 0 \end{cases}, \quad (3)$$

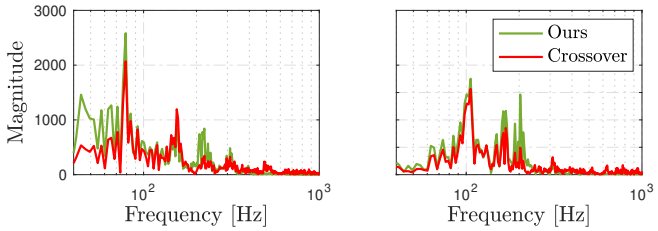


Fig. 3. Examples of the Fast Fourier Transform (FFT) magnitude of the difference between enhanced and input frames for the proposed ("Ours") and crossover network ("Crossover") methods.

which combines in a piecewise fashion the two NLDs: when the input is positive, the function follows $\mathrm{atsr}(x)$; conversely, when the input is negative, $f_{\mathrm{PW}}(x)$ follows $\tanh(2.25x)$. The compression of the function $\tanh(x)$ by means of coefficient 2.25 is necessary for guaranteeing the continuity of $f'_{\mathrm{PW}}(x)$, i.e., the first derivative of $f_{\mathrm{PW}}(x)$. The proposed NLD is thus designed to comply with all of the aforementioned requirements. Finally, the characteristic curve of the proposed nonlinear device (in black) together with $\tanh(2.25x)$ (in orange) and $\mathrm{atsr}(x)$ (in blue) are represented in Fig. 2.

It is worth adding that processing one stem at a time not only prevents inharmonic distortion but also allows us to employ a different NLD for each single stem, opening up new possibilities as far as harmonic generation is concerned. In our implementation, however, we set $f_1 = f_2 = f_{\mathrm{PW}}$.

C. Normalization and Output Stage

Depending on the amplitude of the input signals, the NLDs risk being mostly visited in their quasi-linear regions, severely impairing the generation of harmonics [18]. This is crucially relevant since different MDMs and filters could produce stems with significantly different energy. To sort this issue out, we introduce the normalization functions $\varphi_1, \ldots, \varphi_P$. Their aim is to scale the NLD inputs in $[-1, 1]$ regardless of their amplitude, such that the nonlinearity is fully visited. Many possible normalization functions can be considered. For instance, we may define such functions as follows

$$\varphi_p(\mathbf{x}_p) := \frac{\beta_p}{\max\left\{\left|\mathbf{x}_p[j]\right|\right\}_{j=k}^{k+K-1} + \varepsilon}\,\mathbf{x}_p\,, \quad p = 1, \ldots, P\,, \tag{4}$$

where ε is a small scalar introduced to avoid numerical problems, whereas $\beta_p \in (0, 1]$ is a real-valued hyperparameter introduced for further tuning the generation of overtones.

Finally, we can write the enhanced audio frame $\mathbf{x}_{\mathrm{e}}$ referred to our implementation as follows

$$\mathbf{x}_{\mathrm{e}} = \mathbf{x} - \mathbf{x}_1 - \mathbf{x}_2 + w_1\,\mathbf{f}_1(\varphi(\mathbf{x}_1)) + w_2\,\mathbf{f}_2(\varphi(\mathbf{x}_2))\,, \tag{5}$$

where w_1 and w_2 are two real-valued coefficients that properly weight the harmonically-enriched stems. Notice that w_p and β_p are free parameters that may be varied to tune the algorithm according to, e.g., input track, demixing model, NLD, and transducer under consideration.

Fig. 3a shows the Fast Fourier Transform (FFT) magnitude of the difference between $\mathbf{x}_{\mathrm{e}}$ and $\mathbf{x}$ for the specific case of 12 s from "By The Way" by the Red Hot Chili Peppers, while Fig. 3b from "Get up (I Feel Like Being A) Sex Machine" by James Brown. Such audio tracks are also part of the perceptual test that will be described in Section III. The green and red lines represent the results of the proposed method and crossover network (one of the baselines considered in Section III), respectively. The proposed processing chain is able to generate new harmonics in the low region of the spectrum without compromising high-end frequency content. Moreover, with respect to the crossover network, our method generates higher amplitude harmonics in the mid-frequency range of the spectrum, i.e., those responsible for the missing fundamental effect.

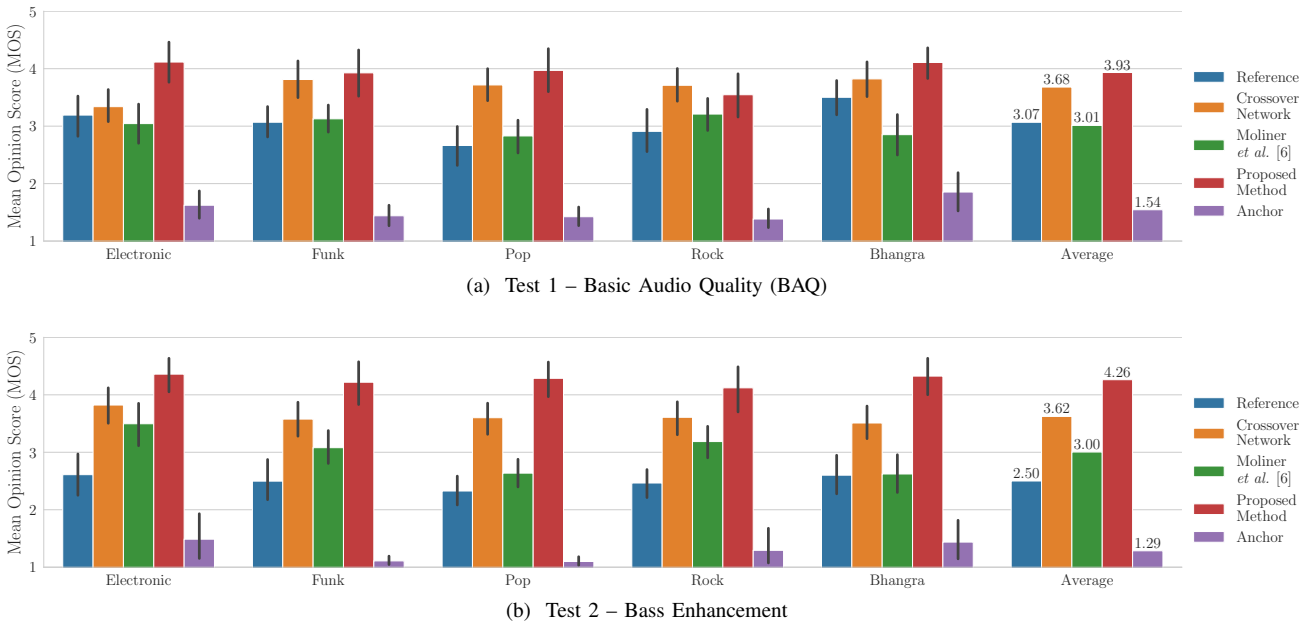
TABLE I
SONGS INCLUDED IN THE PERCEPTUAL TEST.

Song	Artist	Music Genre
Giorgio, By Moroder	Daft Punk	Electronic
Get up (I Feel Like Being A) Sex Machine	James Brown	Funk
Oops!... I Did It Again	Britney Spears	Pop
By The Way	Red Hot Chili Peppers	Rock
Mundian To Bach Ke	Punjabi MC	Bhangra

III. Perceptual Test

In order to compare the proposed method with other techniques available in the literature, we run an experimental campaign employing a web-based software compliant with different ITU protocols [19]. In particular, we modify the tool for performing a Mean Opinion Score (MOS) test (ITU-T Rec. P.800.1.) The listeners were asked first to rate excerpts (12 s long) of the songs listed in Table I on a scale of 1 (Bad) to 5 (Excellent) as far as Basic Audio Quality (BAQ) is concerned (Test 1). Then, they were asked to rate the very same tracks on the same scale as far as bass enhancement is concerned (Test 2). Specifically, Test 2 is presented to the subjects only after completing Test 1.

The five audio tracks were processed by means of three different algorithms: the method proposed in this manuscript, a time-domain method based on a crossover network similar to the one shown in [5] (later on referred to just as "crossover network"), and the hybrid method proposed in [6], which represents the state-of-the-art technique. All the algorithms were written in Python, apart from the hybrid method for which we avail of the MATLAB code released by the authors [6]. In particular, as for the implementation of our method, β_p takes value in $\{0.8, 0.9\}$, whilst w_p takes value in $\{1, 1.5, 2\}$, according to the specific stem of each song. For a fairer comparison with the crossover network, instead, we decided to apply the same weight w_1 and the same NLD employed for processing the `bass` stem (i.e., f_{PW}) considered for our method. While the two time-domain methods (i.e., the proposed method and the crossover network) were able to perform stereo processing in real-time on CPU, the hybrid method in [6] was characterized by a processing time that prevented it from reaching real-time execution.



(a) Test 1 – Basic Audio Quality (BAQ)

(b) Test 2 – Bass Enhancement

Fig. 4. Results of the perceptual test with 95% confidence intervals. (a) Scores related to the five different music genres as far as Basic Audio Quality (BAQ) is concerned; (b) scores related to the five different music genres as far as bass enhancement is concerned. For both tests, the rightmost bar chart reports the average scores.

The three processed tracks together with the reference track (later on referred to as "Reference") were highpassed at 200 Hz [6]. On the one hand, this was to simulate the cut-off of a small-size loudspeaker regardless of the frequency response of the listening devices used by the assessors during the test. On the other hand, this allows us to assess just the perception of low tones given by the "missing fundamental" phenomenon. In addition, the listeners were asked to rate also the highpassed version of the reference track at 500 Hz (later on referred to as "Anchor"), i.e., the baseline of the perceptual scale. In order to avoid biases due to the perceived loudness, we applied to all the items first a peak normalization at -1.0 dB and then a LUFS (Loudness Unit referenced to Full Scale) normalization at -14.0 dB LUFS (i.e., Spotify's setting recommendations.) Such normalizations, whose implementations are compliant with ITU-R Rec. BS.1770-4 [20], were applied to the high-passed tracks. Audio examples are available online.[1]

Altogether, 30 people participated to the perceptual test (with an average age of 26 years.) They rated five different versions of five songs twice: first for Test 1 (BAQ) and then for Test 2 (bass enhancement). They took the two tests on consumer-grade headphones, were trained to recognize "good" from "bad" tracks for both tasks, and completed the overall perceptual test in about 20 minutes.

Figure 4 shows the results of Test 1 and Test 2 with the 95% confidence intervals and the relative average scores (rightmost bar charts.) In particular, as far as Test 1 shown in Fig. 4a is concerned, the assessors rated the BAQ quality of the proposed method always greater than the Reference's one and as the highest for all the music genres but Rock, for which the crossover network received a higher score. The excerpt related to Rock genre was, in fact, characterized by a very dense bass line, and thus the listeners may have deemed the bass enhancement provided by the proposed method a little excessive. Nevertheless, as pointed out in Fig. 4a, the proposed technique received, on average, the score 3.93, against, e.g., the score 3.68 of the crossover network, performing better than the other algorithms. Instead, as far as Test 2 shown in Fig. 4b is concerned, the listeners rated the proposed method as the best method for enhancing the perception of low frequencies in all music genres, receiving on average the score of 4.26, against the 3.62 of the crossover network, i.e., the best baseline.

We can thus conclude that the proposed algorithm is able to outperform other algorithms available in the literature, including state-of-the-art methods, in terms of both BAQ and bass enhancement, being at the same time characterized by a low computational cost enabling inference approximately three times faster than real-time.

IV. Conclusions

In this letter, we proposed a new time-domain method for enhancing the perception of low tones in music applications. In particular, the method is characterized by a novel processing pipeline, as well as by a new nonlinear device, and takes advantage, for the first time, of deep-learning-based Music Demixing Models. MDMs are exploited to prevent the generation of inharmonic distortion, i.e., the main drawback of traditional time-domain methods. Through an experimental campaign, we compared the method to other techniques available in the literature. The proposed algorithm outperformed the state-of-the-art in both basic audio quality and bass enhancement for different music genres, still being able to process stereo tracks in real-time on CPU. In the future, thanks to its general formulation, the algorithm will be able to accommodate more sophisticated MDMs, improving its performance even further.

[1][Online]. Available: https://polimi-ispl.github.io/vbe-demixing/

REFERENCES

[1] E. Zwicker and H. Fastl, *Psychoacoustics - Facts and Models*. Springer Berlin Heidelberg, 1999, vol. 22.

[2] H. Musmann, "Genesis of the mp3 audio coding standard," *IEEE Transactions on Consumer Electronics*, vol. 52, pp. 1043–1049, 8 2006.

[3] E. Larsen and R. M. Aarts, *Audio Bandwidth Extension: Application of Psychoacoustics, Signal Processing and Loudspeaker Design*. John Wiley and Sons, Ltd, 2004.

[4] D. Ben-Tzur, "The effect of the maxxbass 1 psychoacoustic bass enhancement system on loudspeaker design," in *Proceedings of the 106th Audio Engineering Society Convention*, 5 1999.

[5] N. Oo, W.-S. Gan, and M. O. J. Hawksford, "Perceptually-motivated objective grading of nonlinear processing in virtual-bass systems," *Journal of the Audio Engineering Society*, vol. 59, pp. 804–824, 12 2011.

[6] E. Moliner, J. Rämö, and V. Välimäki, "Virtual bass system with fuzzy separation of tones and transients," in *Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx2020)*, 9 2020.

[7] N. Oo and W.-S. Gan, "Harmonic analysis of nonlinear devices for virtual bass system," in *Proc. Int. Conf. Audio, Language, and Image Processing*, 8 2008, pp. 279–284.

[8] L. K. Chiu, D. V. Anderson, and B. Hoomes, "Audio output enhancement algorithms for piezoelectric loudspeakers," in *Proceedings of the Digital Signal Processing and Signal Processing Education Meeting (DSP/SPE)*, 1 2011, pp. 317–320.

[9] M. R. Bai and W.-C. Lin, "Synthesis and implementation of virtual bass system with a phase-vocoder approach," *Journal of the Audio Engineering Society*, vol. 54, pp. 1077–1091, 2006.

[10] T. Lee, S. Lee, Y. cheol Park, and D. H. Youn, "Virtual bass system based on a multiband harmonic generation," in *2013 IEEE International Conference on Consumer Electronics (ICCE)*, 1 2013, pp. 399–400.

[11] A. J. Hill and M. O. J. Hawksford, "A hybrid virtual bass system for optimized steady-state and transient performance," in *Proceedings of the 2nd Computer Science and Electronic Engineering Conference (CEEC)*, 9 2010, pp. 1–6.

[12] P. Hoffmann and B. Kostek, "Bass enhancement settings in portable devices based on music genre recognition," *Journal of the Audio Engineering Society*, vol. 63, pp. 980–989, 1 2016.

[13] R. Hennequin, A. Khlif, F. Voituret, and M. Moussallam, "Spleeter: A fast and efficient music source separation tool with pre-trained models," *Journal of Open Source Software*, vol. 5, p. 2154, 6 2020.

[14] A. Défossez, "Hybrid spectrogram and waveform source separation," in *Proceedings of the Music Demixing Workshop*, 11 2021.

[15] M. Kim, W. Choi, J. Chung, D. Lee, and S. Jung, "Kuielab-mdx-net: A two-stream neural network for music demixing," in *Proceedings of the Music Demixing Workshop*, 11 2021.

[16] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proceedins of the 18th International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 10 2015, pp. 234–241.

[17] S. K. Mitra, *Digital Signal Processing*, 2nd ed. New York: McGraw-Hill, 2001.

[18] R. Giampiccolo, A. Bernardini, and A. Sarti, "A time-domain virtual bass enhancement circuital model for real-time music applications," in *2022 IEEE 24th International Workshop on Multimedia Signal Processing (MMSP)*, Shanghai, China, 2022, pp. 1–5.

[19] M. Schoeffler, S. Bartoschek, F. R. Stöter, M. Roess, S. Westphal, B. Edler, and J. Herre, "webmushra - a comprehensive framework for web-based listening tests," *Journal of Open Research Software*, vol. 6, 2018.

[20] C. J. Steinmetz and J. D. Reiss, "pyloudnorm: A simple yet flexible loudness meter in python," in *Proceedings of the 150th AES Convention*, 4 2021.