Audio Engineering Society

Convention Paper 10648

Presented at the 154th Convention
2023 May 13-15, Espoo, Helsinki, Finland

*This paper was peer-reviewed as a complete manuscript for presentation at this convention. This paper is available in the AES E-Library (http://www.aes.org/e-lib) all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.*

# Advances in Perceptual Bass Extension for Music and Cinematic Content

Sunil Bharitkar, Ema Souza Blanes, Glenn S. Kubota, and Ashish Rawat

*Samsung Research America, DMS Audio, Valencia CA, USA*

Correspondence should be addressed to Sunil Bharitkar (`s.bharitkar@samsung.com`)

## ABSTRACT

Small form-factor and thin devices exhibit a high-pass frequency response due to loudspeaker-enclosure constraints. The low-frequency reproduction loss from these devices severely degrades the audio experience for music and cinematic content. This paper presents a new perceptual bass extension model using a side chain for music and cinematic content and leveraging the principle of the missing fundamental frequency. Optimizing the nonlinear function parameters enables the nonlinear function output to be invariant to input signal level changes. The model employs a unique input gain normalization scheme based on loudness metadata and level-matching between multiple side chains. A loudness compensation algorithm restores the perception of bass, particularly at low playback levels. Subjective testing and perceptually derived objective metrics using television (TV) loudspeakers validate the performance of the approach.

## 1 Introduction

There is a strong interest in psychoacoustic bass enhancement to improve the perceived bass frequencies in audio signals from thin loudspeakers in small form-factor enclosures. The traditional techniques involve leveraging the percept of the missing fundamental where the pitch of the fundamental is perceived in the difference-frequencies of harmonics if the fundamental is incapable of being reproduced by the loudspeaker [1], [2], [3]. Conventionally [4] a side-chain synthesizes the harmonic coefficients from the input audio signal using a nonlinear function and a gain weighting term. The side-chain signal is mixed in with a delayed and high-pass filtered version of the input signal before delivering the output to the loudspeaker. A technique for generating both even and odd harmonics, using a half-wave rectifier for synthesizing even harmonics and a clipper for odd harmonics, is presented in [5]. The side-chain included downsampling and upsampling operations in conjunction with short-time Fourier transform for multi-band even and odd harmonic synthesis. Shi *et al.* [6] demonstrated successful subjective test results using a mapping of two nonlinear functions (arc-tangent and exponent) to create the perception of bass on parametric loudspeakers. Oo [7] analyzes the arc-tangent-based nonlinear function and derives a closed-form expression for the amplitude and phase of individual harmonics. Giampiccolo [8] presents an analog circuit model comparing three nonlinear functions (including exponential nonlinearity) for real-time music applications. The subjective testing results indicate that different nonlinear functions are better suited to different types of music (pop, electronic, R&B). A frequency-domain Prony's method to estimate the parameters of a low-pass input signal before harmonic synthesis is presented in [9]. Hoffmann *et al.* [10] present a music genre-classifier with a nonlinear function (e.g., exponential) based harmonics synthesis for virtual bass, and compare results with MaxxBass[1]. An adaptive frequency tracking method [11] using analysis filterbanks and Linear Predictive Coding (LPC) to generate harmonics of the dominant fundamental frequency. Subsequently, a synthesis bandpass filter (after the nonlinear function) is adjusted to select the harmonics. A perceptually motivated objective grading system [12], using the *Rnonlin* distortion model [13], is used to identify classes of *good* (e.g., exponential nonlinearity), *bass-killer*, *not-recommended*, and *highly distorted* nonlinear functions.

On the other hand, a spectrum-shifting phase-vocoder [14] elicits an improved bass preference than the hidden reference but consistently showed *worse* audio quality preference (in terms of noise and distortion) compared to the reference (cf. Fig. 15-17 in [14]). Nonetheless, as shown in their paper, the phase-vocoder compares better in both bass and audio quality with the state-of-the-art MaxxBass system that employs a nonlinear function. Hybrid approaches (e.g., [15], [16], [17]) combining nonlinear function with a phase vocoder attempt to leverage the benefits of both models for transient/percussive and harmonic (steady-state) signals. However, such approaches are compute-intensive and introduce latency constraints for real-time processing, especially with associated video. Mu [18] *et al.* compare harmonic weighting schemes while tracking the fundamental frequency using subjective tests and a perceptually-derived objective metric. According to the authors, the frequency-dependent tracking and frequency-based harmonic synthesis replace a time-domain nonlinear function since the time-domain nonlinear function is input signal level dependent (viz., the harmonics' amplitude and the harmonics' spectrum envelope are input signal level-dependent). The perceptual metric was premised on Model Output Variables (MOV) using Perceptual Evaluation of Audio Quality (PEAQ) [19].

In this paper, we advance the conventional approach (state-of-the-art) with a model (i) employing a nonlinear function that is tuned to create controlled harmonics where the harmonic excitation pattern (amplitudes and envelope) is invariant to input signal level, (ii) having a metadata-driven input gain normalization scheme, (ii) including a level-matching mechanism between the low-frequency effects and L+R side-chains, (iii) incorporating a loudness compensation scheme using International Standards Organization (ISO) 226 [20] loudness contours[2] to maintain the perception of bass at low playback levels, and (iv) tested on music and cinematic content. We also perform exploratory analysis between preference listening tests and multiple perceptually derived objective metrics, including *Rnonlin* [13] and PEMO-Q [21]. Section 2 presents the proposed model and associated optimizations for a Digital Signal Processor (DSP) implementation. Section 3 provides the subjective test results and ties these results with objective metrics. Section 4 concludes/summarizes the paper.

## 2 The Perceptual Bass Extension Model

Fig. 1 shows the model for perceptual bass extension using the proposed time-domain nonlinearity with a loudness metadata parser. A heuristic classifier based on the number of audio channels in the embedded metadata, associated with the encoded audio or video format, determines the content class (e.g., cinematic-movies, documentaries) or non-cinematic[3].

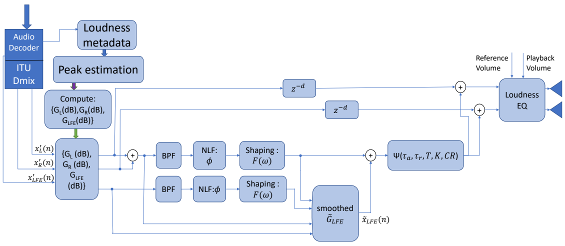


**Fig. 1:** The perceptual bass extension model.

### 2.1 ITU Downmix with LFE

Only in the case of cinematic content, the preprocessing step involves an ITU downmix [22] of the left $x_L(n)$, right $x_R(n)$, surrounds ($x_{LS}(n)$, $x_{RS}(n)$) and center $x_C(n)$ channels as depicted in Fig. 1 at the output of the audio decoder. In the case of music signals, the low-frequency effect (LFE) side-chain is disabled since the LFE channel is not present. The LFE channel

[1] Waves Audio, `https://www.waves.com/`

[2] A new version of the ISO document on loudness contours is under development and due to be published in 2023

[3] The non-cinematic content is assumed to be music

$x_{LFE}(n)$ undergoes harmonic synthesis processing for cinematic content.

$$\begin{aligned} x'_L(n) &= x_L(n) + 0.707x_C(n) + 0.707x_{LS}(n) \\ x'_R(n) &= x_R(n) + 0.707x_C(n) + 0.707x_{RS}(n) \\ x'_{LFE}(n) &= x_{LFE}(n) \end{aligned} \tag{1}$$

The side chain comprises two branches in the case of cinematic content (i) a mono-downmix for $x'_L(n)$ and $x'_R(n)$ input to a band-pass filter (BPF) that selects the portion of the input signal that the loudspeakers cannot reproduce, and (ii) the $x'_{LFE}(n)$ input to the second band-pass filter (BPF). In the case of music content, the LFE side-chain is disabled.

## 2.2 Metadata-based Input Gain Control

Streaming media include metadata either in the container or the audio codec indicating the loudness information $Loud_i$ $(i = \{L, R\})$ for the full file. The presented model leverages this loudness metadata to normalize the input signal gain to a reference value $R$ (dB) below 0 dBFS. Towards this goal, multiple models (below) were evaluated to predict the gain, $P_i$ $(i = \{L, R\})$ required for the peak in the music file to attain 0 dBFS using the $Loud_i$ loudness metadata. The input normalization to a reference gain $R$ (dB) is performed to minimize any audible effects from the interaction between the nonlinear compressor (described later) with the combined time-domain harmonic-synthesis nonlinear function $\phi(.)$ and the harmonic shaping filter $F(\omega)$. An alternative approach would be to normalize to a reference based on frame-level analysis, but this would introduce audible artifacts, such as gain fluctuations between frames. The gains $G_i$ $(i = \{L, R\})$ are expressed as,

$$G_i = -R + P_i; \qquad (i = \{L, R\}) \tag{2}$$

In the case of file-based media, the file-based loudness metadata resides in the header for stereo content. A metadata parser extracts this loudness metadata (in hexadecimal format) and converts it to a decimal number. To develop a model $f_i$ that maps from loudness Loud$_i$ space to $P_i$, a dataset of over 200 music files was analyzed by parsing the metadata and also by downloading the corresponding music file, decoding to *wav* format, and then computing $P_i$. Three candidate models, including least-squares, regression, and neural network, were assessed to predict $P_i$ from Loud$_i$ using the dataset.

### 2.2.1 Least-squares (Lsq)

The least-squares optimal model $\mathbf{W}^* \in \Re^{2\times 2}$ is obtained by a pseudo-inverse of the matrix $\Lambda \in \Re^{N\times 2}$ comprising the loudness metadata converted to decimal and post-multiplying with the matrix $\Phi \in \Re^{N\times 2}$ comprising the amount required in dB for the peak amplitude in the content to reach 0 dBFS, and $N$ is the number of content-files used for developing the model. We used 80% of the 202 files for the least squares model development and 20% to test performance. Specifically,

$$\Lambda = \begin{bmatrix} Loud_L^{(1)} & Loud_R^{(1)} \\ Loud_L^{(2)} & Loud_R^{(2)} \\ \cdot & \cdot \\ Loud_L^{(N)} & Loud_R^{(N)} \end{bmatrix} \quad \Phi = \begin{bmatrix} P_L^{(1)} & P_R^{(1)} \\ P_L^{(2)} & P_R^{(2)} \\ \cdot & \cdot \\ P_L^{(N)} & P_R^{(N)} \end{bmatrix} \tag{3}$$

Accordingly,

$$\begin{aligned} \Phi &= \Lambda \mathbf{W} \\ \mathbf{W}^* &= (\Lambda^T \Lambda)^{-1} \Lambda^T \Phi \end{aligned} \tag{4}$$

### 2.2.2 Linear Regression (Lreg)

The linear regression model employs quadratic terms to determine the $(\alpha_i, \beta_i)$ coefficients and is represented as,

$$\begin{aligned} P_L &= \alpha_0 + \alpha_1 Loud_L + \alpha_2 Loud_R + \alpha_3 Loud_L^2 \\ &+ \alpha_4 Loud_R^2 \qquad (5) \\ P_R &= \beta_0 + \beta_1 Loud_L + \beta_2 Loud_R + \beta_3 Loud_L^2 \\ &+ \beta_4 Loud_R^2 \end{aligned}$$

The unweighted (ordinary least-squares) fit yields the best results for the present dataset after comparing different weighting schemes (e.g., logistic, Huber, Andrews, Cauchy, etc.).

### 2.2.3 Nonlinear Regression (Nreg)

The nonlinear regression model employs quadratic terms to determine the $(\delta_i, \gamma_i)$ coefficients and is represented as,

$$\begin{aligned} P_L &= \delta_0 + \delta_1 Loud_L^{\delta_2} + \delta_3 Loud_R^{\delta_4} \\ P_R &= \gamma_0 + \gamma_1 Loud_L^{\gamma_2} + \gamma_3 Loud_R^{\gamma_4} \end{aligned} \tag{6}$$

For developing this model, the dataset size was 80% of the 202 files that were scraped for loudness metadata and correspondingly downloaded and decoded to wav format.

#### 2.2.4 Neural network (FCNN)

A fully-connected feed-forward neural network [23] was optimized using 60% training and 20% validation dataset size ($N = 202$ files) after performing *Bayesian optimization* [24] by using the number of layers and the number of neurons per layer with the hyper-parameters. The final network had one hidden layer with seven neurons, *tanh* activation, and a linear output layer.

#### 2.2.5 Comparative Results

Fig. 2 shows the results for estimating $P_L$ and $P_R$ from the loudness metadata for both channels on the test set of size 34 (total 68 samples for both channels). The neural-network model gives the best result, which is validated in the Table below after using the mean-square-error between the true $P$ and the predicted value $\hat{P}$

$$MSE = \frac{1}{34}\left(\sum_{i=1}^{34}(P_L^{(i)} - \hat{P}_L^{(i)})^2 + (P_R^{(i)} - \hat{P}_R^{(i)})^2)\right) \quad (7)$$

| | Lsq | Lreg | Nreg | FCNN |
|---|---|---|---|---|
| MSE | 4.38 | 0.18 | 0.23 | 0.13 |

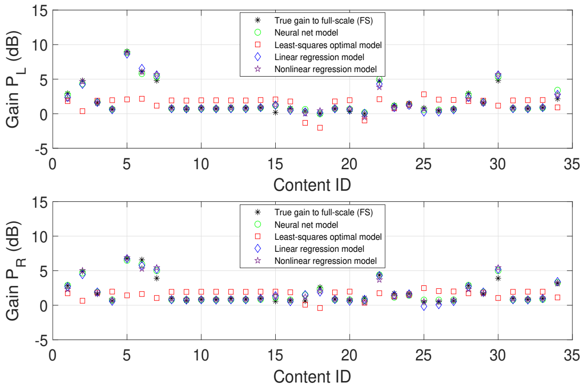


**Fig. 2:** Modeling loudness metadata to a full-scale gain of the maximum audio sample value for the given content.

### 2.3 Nonlinear Function (NLF)

Based on [4] and [7], a nonlinear function that *necessarily* generates both even and odd harmonics is appropriate for perceptual bass enhancement. However, there is no guarantee that this class of functions will be *harmonically stable* as a function of the fundamental frequency and the level of the input signal ([7] and as shown and explained related to Fig. 14b). Accordingly, by properly tuning the NLF with two tuning parameters $\kappa_1$ and $\kappa_2$, the harmonic excitation pattern (number of harmonics and the envelope of the harmonics) remains invariant to frequency and input signal level,

$$y(n) = \begin{cases} \kappa_1(x(n) + (1-\kappa_1)y(n-1)); x(n) > y(n-1) \\ \kappa_2 x(n) + (1-\kappa_2)y(n-1); else \end{cases} \quad (8)$$

An example of a properly tuned NLF (8) for $f_0 = 45$ Hz, generating even and odd harmonics, is shown in Fig. 3. The filtered harmonics, depicted in the black curve, is the output after a shaping response filter $F(\omega)$, which suppresses the fundamental and attenuates harmonics beyond the second harmonic $f_2$. In this work, only the first two harmonics, $f_1$ and $f_2$ of the fundamental $f_0$, are used to minimize the impact of intermodulation components. Additionally, the shaping response filter, $F(\omega)$, was fixed to a cascade of a band-pass filter to pass the second and third harmonics and a second-order biquad filter in the pass-band. The harmonic pattern is invariant in terms of the number of harmonics and the shape and envelope of the harmonic spectrum, as is evident in the orange curve in Fig. 4 for a sinusoid with $f_0 = 50$ Hz (green curve) at input levels of -20 dB, -10 dB, 0 dB, and 10 dB. Comparatively, the *good* exponential NLF [7] $(e - e^{1-x})/(e-1)$ shows harmonic instability (blue curve) and which is a common concern with time-domain nonlinear functions. By performing a sinusoidal signal sweep below a loudspeaker cutoff frequency (viz., in this setup being 75 Hz) and comput-

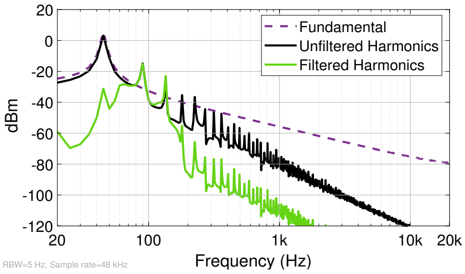
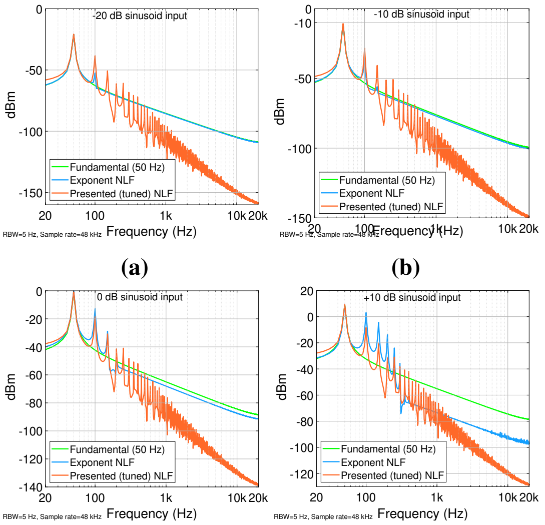


**Fig. 3:** Example of shaped harmonic response for $f_1$ and $f_2$ with $f_0 = 45\ Hz$ (Note: The harmonic shaping also rejects the fundamental frequency).



**(a)** **(b)**

**(c)** **(d)**

**Fig. 4:** Influence of level on the harmonic pattern of a sinusoidal signal of 50 Hz (blue: exponential NLF, orange: presented NLF, green: fundamental) (a) -20 dB, (b) -10 dB, (c) 0 dB, (d) 10 dB.

ing the level difference between the first two desired harmonics $f_1$ and $f_2$ (with fundamental $f_0$) at various input levels, one can observe the robustness of the harmonic excitation pattern in Fig. 5. Specifically, the x-axis is the sinusoid input frequency $f_0$ and the y-axis is the level difference, $|f_1 - f_2|$ between the first two harmonics that are delivered to the shaping filter $F(\omega)$. As can be seen, the presented NLF exhibits all the desirable properties, such as generating both even and odd harmonics. The harmonic pattern also maintains invariance to the fundamental frequency and the input signal level.

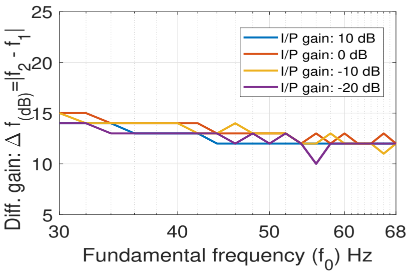


**Fig. 5:** Difference between the first two harmonics as a function of the fundamental frequency for different input sinusoidal signal gain.

## 2.4 Side-chain Level Matching

After NLF, the level difference between the LFE and the L+R mono-downmix side-chains for cinematic content is matched to the level difference between both side-chains before the NLF to maintain the relative loudness balance arising from the individual side-chain processing. Fig. 1 depicts the processing in the side-chain with the details shown in Fig. 6. The signals with superscript $(f)$ and $(h)$ represent the input to the NLF and the output from the NLF, respectively. The dB gain between these two side chains is compared before and after NLF processing, and a correction gain, as the output from a first-order low-pass smoother, is applied to the NLF-processed LFE side chain. Example time-domain signals are shown in Fig. 7

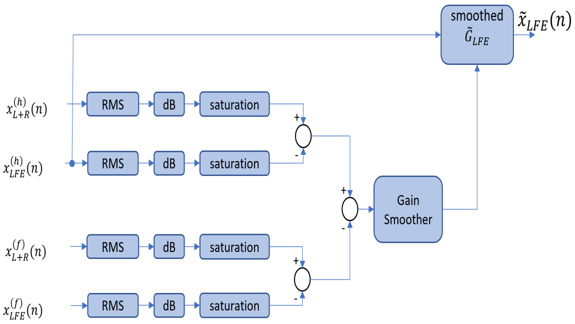
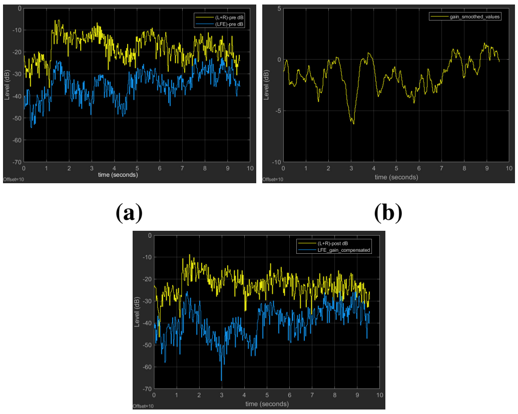


**Fig. 6:** The signal level normalization for the low-frequency effects channel.

## 2.5 Compressor

The input gain normalization using the predicted peak amplitude keeps the side-chain compressor, $\Psi(\tau_a, \tau_r, T, K, CR)$, contribution to a minimum. This minimal interaction with the NLF is achieved using small values for the compression ratio $CR = 1.25$, threshold $T = -1.5$ (dB), and knee width $K = 0.5$ dB. The attack time $\tau_a$ and release time $\tau_r$ constants being 0.01 (sec) and 0.15 (sec), respectively. The core equation governing the compressor core output signal,



**(a)** **(b)**

**(c)**

**Fig. 7:** Side-chain signals from Batman: The Dark Knight (yellow: L+R mono, blue: LFE) (a) before level matching before NLF, (b) smoothed LFE gain, (c) after level-matching and after NLF.

$x_{dB}^{(comp)}(n)$, is,

$$x_{dB}^{(comp)}(n) = \begin{cases} x_{dB}(n) & x_{dB}(n) < \rho_1 \\ x_{dB}(n) + \Gamma(n) & \rho_1 \leq x_{dB}(n) \leq \rho_2 \\ T + \frac{(x_{dB}(n) - T)}{CR} & x_{dB}(n) > \rho_2 \end{cases} \tag{9}$$

where, $x_{dB}(n) = 20\log_{10}|x(n)|$, $\Gamma(n) = ((1/CR) - 1)(x_{dB}(n) - T + (K/2))^2/2K$, $\rho_1 = (T - \frac{K}{2})$, and $\rho_2 = (T + \frac{K}{2})$. The compressor core output is applied to a low-pass filter with the aforementioned attack and release time constants. The resulting signal is mixed in with the original signal filtered by a constant delay.

### 2.6 Playback Loudness Compensation (Dynamic EQ)

ISO 226 [20] has introduced updated equal-loudness contours for different levels from 20 dB through 100 dB (referenced at 1 kHz). The contours show that the human auditory response spectrum changes with level, leading to a loss of bass and higher frequencies at low levels. The TV speakers' C-weighted sound pressure level (SPL) was measured using calibrated pink noise at different playback volume settings, with the SPL mapped to the corresponding ISO curve. For a reference volume setting of 60 (mid-volume reference level), the SPL was 73 dBC. For lower volume (volume setting 30 on the TV), the SPL mapped to 65 dBC, whereas for the high volume condition (volume setting 80), the SPL was 77 dBC. Differential loudness contours [25] were computed between the reference level ISO curve and the low and high volume ISO curves. Fig. 8 depicts an example of this process for the reference mid-volume curve (blue) and low-volume (red) curves normalized at 1 kHz. The low-frequency and high-frequency gains (shown in green) are computed based on these curves' differences. A cascade of shelf filters models this difference with an additional high shelf, ensuring no gain below the cutoff frequency of the TV speakers (≈ 75 Hz). The purple curve shows the resulting low-volume compensating filter.

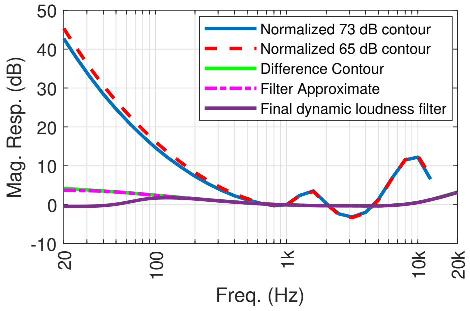


**Fig. 8:** Example loudness compensating filter between reference playback volume and low volume.

### 2.7 Test Platform

The optimized PBE algorithm, leveraging NLF compute optimizations, was tested on a custom board with a single-core ARM Cortex-M7 processor running at 600 MHz. The input and output interfaces were S/PDIF, running at 48 kHz. This processor has a floating-point math unit; hence, the C/C++ code used floating-point throughout. Some speed improvement may be realized if fixed-point processing were used. The performance of the stereo floating-point PBE algorithm on the ARM M7 is 5.6 MCPS. For a stereo + LFE input, the PBE algorithm uses 26.7 MCPS.

## 3 Results

Fig. 9 (blue curve) shows the resulting reference-volume pink-noise spectrum simulated at the input of the TV loudspeakers with the processing disabled, whereas the black curve shows the resulting pink-noise spectrum with the perceptual processing. The relative curves show that the NLF and shaping filter response $F(\omega)$ introduce a negligible change in the pink-noise spectrum from the off condition. The measurements at the output of the TV speakers will show additional roll-off below the cutoff frequency of $\approx$ 75 Hz.

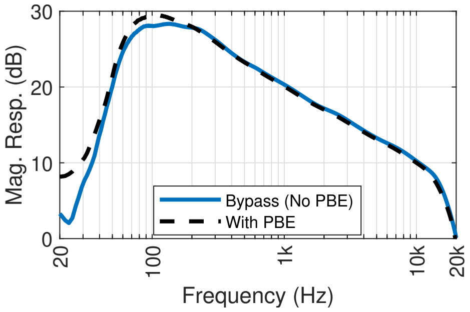


**Fig. 9:** Pink noise spectrum comparing the proposed approach versus reference (PBE bypass).

### 3.1 Subjective Testing

A listening test was conducted on a properly equalized and level-calibrated television (Samsung GQ55 QN90 TV) in a listening room to assess the effect of the presented algorithm on the listener's preference. Stereo audio samples were processed using the PBE and PBE with Dynamic EQ algorithms. The test consisted of a comparison between the Reference content, PBE, and PBE with Dynamic EQ. As the dynamic algorithm adjusts the PBE processing to the playback volume, the test was divided into three sessions, each corresponding to one playback level. Specifically, (i) medium level (Volume 60) for which the PBE and PBE+dynamic EQ conditions are the same (73 dBC), (ii) higher level (Volume 80) at 77 dBC, and (iii) lower level (Volume 30) at 65 dBC. Three cinematic and music tracks were selected, as detailed in Table 1, which resulted in six trials for each session. The assessors listened to the three listening conditions, A, B, and C, and rated them according to their preference on a scale ranging from 0 to 100. A Max/MSP custom interface enabled the listeners to control the playback and to give their scores. The arrangement of the conditions and the order of presentation of the audio samples were automatically randomized. The three test sessions were randomly assigned between subjects. Thirteen assessors participated in each listening test session, including eight trained and five naive listeners. All three sessions were analyzed independently. The results for the low, medium, and high levels are shown in Figs. 11, 12, and 13 respectively. The data were screened for outliers and tested for homoscedasticity and normal distribution. A 2-way ANOVA procedure was then employed to evaluate the effect of the Condition and Sample factors on scores, and paired t-tests were applied to compare the mean scores of each pair of conditions.

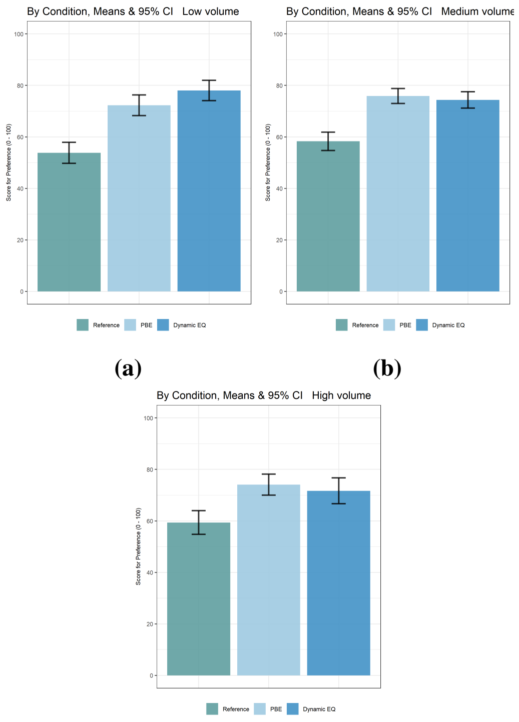
**Table 1:** Listening material

| Label | Type | Content | Source | Duration |
|---|---|---|---|---|
| TDK | Cinematic | Bank robbery scene | Warner Bros. *The Dark Knight* | 11s |
| INT | Cinematic | Music and dialog | Paramount Pictures *Interstellar* | 13s |
| MM | Cinematic | Cars engines, gun fights | Warner Bros. *Mad Max: Fury Road* | 20s |
| AF | Music | Electronic music | H. Faltermeyer/ MCA Records *Axel F* | 16s |
| NP | Music | Rock music, vocals and bass guitar | A.R Rahman *Nadaan Parinde* | 16s |
| SA | Music | Disco | Bee Gees/ RSO Records *Stayin'Alive* | 8s |

In all three sessions (Fig. 10), the effect of the listening condition is statistically significant, with no significant effect of the sample factor. The PBE and PBE+Dynamic EQ were rated closely, while the reference sample scored lower (see Table 2). Higher discrimination between the conditions is observable with the low-volume setting. The PBE+Dynamic EQ condition scored significantly higher than the PBE stimuli for the low-volume test, while no significant difference was noticeable between these two conditions at higher volumes. Only the condition factor was significant in

**Table 2:** Mean scores, per volume level and listening condition

| Mean Score | Reference | PBE | Dynamic EQ |
|---|---|---|---|
| Low Volume | 54.2 | 72.7 | 78 |
| Medium Volume | 58.3 | 75.9 | 74.3 |
| High Volume | 60.6 | 75.3 | 72.9 |



**(a)**

**(b)**

**(c)**

**Fig. 10:** Results (green: reference played from TV, cyan: PBE from TV, blue: PBE+dynamic EQ from TV) for the (a) low volume listening test, per condition, (b) medium volume listening test, per condition, and (c) high volume listening test, per condition.

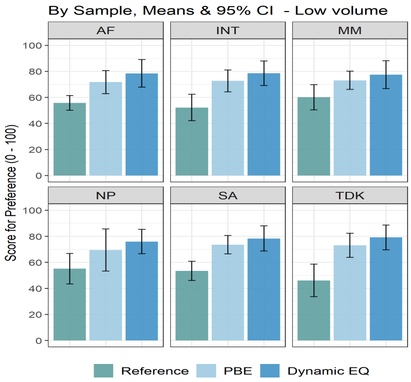


**Fig. 11:** Results for the low volume listening test, per condition and sample

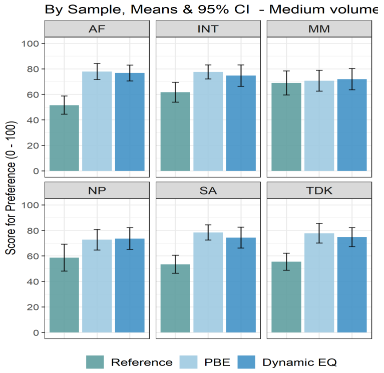


**Fig. 12:** Results for the medium volume listening test, per condition and sample

the lower level test session, with $p_{condition} = 1.1e^{-4}$. Paired comparisons revealed that the difference between all conditions' ratings was statistically significant, with $p_{ref-pbe} = 1.4e^{-10}$, $p_{ref-dyn} = 3.6e^{-11}$ and $p_{pbe-dyn} = 0.0036$. The listening condition factor significantly affected the ratings in the medium-level test, with $p_{condition} = 5e^{-5}$. The PBE and PBE+dynamic EQ conditions received close ratings and did not significantly differ ($p = 0.3$), but both were rated significantly higher than the reference condition ($p_{ref-pbe} = 1.9e^{-12}$, $p_{ref-dyn} = 9.7e^{-11}$). A statistically significant interaction between the effects of the sample and condition factors was found ($p_{interaction} = 1.5e^{-8}$). The MM and NP tracks caused less discrimination between the conditions. The high volume-level test gave similar results, with a significant effect of the condition factor ($p_{condition} = 0.025$). Although the PBE+dynamic EQ condition was rated slightly lower than the PBE condition, the two mean scores did not significantly differ. Like the other test sessions, these two conditions were rated significantly higher than the reference, with $p_{ref-pbe} = 4.6e^{-6}$ and $p_{ref-dyn} = 1.5e^{-4}$. The analysis also showed a significant interaction between the sample and condition factors ($p_{interaction} = 0.003$). A similar rating trend is noticeable for all samples, but some tracks, including MM, INT, and NP, introduced smaller discrimination between the conditions.

### 3.2 Objective Metrics: Exploratory Analysis

Towards developing a perceptual metric suited for perceptual bass extension, an exploratory analysis was performed at reference volume (without dynamic EQ)

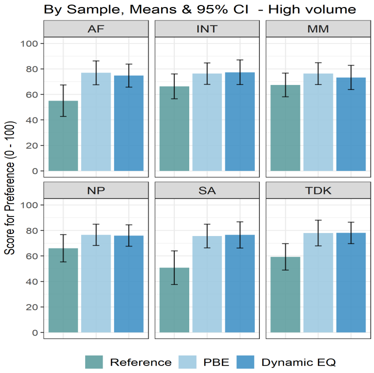
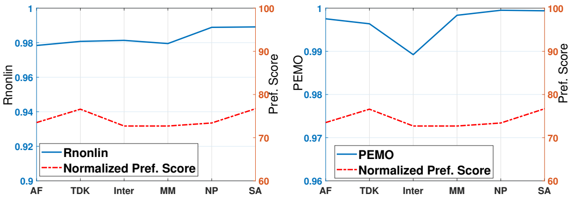


**Fig. 13:** Results for the high volume listening test, per condition and sample

for all test content using perceptual nonlinear distortion metrics *Rnonlin* [12] [7] and PEMO-Q (viz., PSM$_t$) [21]. Both perceptual metrics operate on mono channels, so the output from both channels is averaged to give a single metric for the provided test content. The results from *Rnonlin* and PEMO-Q show that the presented solution attains a high value for all test content, and the trend for these metrics and the preference scores from subjective tests match reasonably well. Furthermore, as an exercise in determining equivalency between the *Rnonlin* values in this paper and the results shown in [7] (cf. Fig. 9 and Table 8 with eq. (20)) for various side-chain NLFs, the *Rnonlin* values using the presented NLF and side-chain map to Mean Opinion Scores (MOS) ranging between 7 and 9.5 for the reference mid-volume test content.



**(a)** **(b)**

**Fig. 14:** Perceptual/objective metrics relative to Preference scores for the 6-test items at medium volume (a) *Rnonlin* and Preference, (b) PEMO-Q and Preference Scores.

## 4 Conclusions and Future Directions

Uncontrollable harmonics are a pervasive problem with time-domain nonlinear functions (NLF). However, the advantages of using time-domain NLFs include low-complexity implementation. This paper presented an NLF selectively tuned to generate harmonics invariant to the input frequency and signal level (viz., controlled harmonics). In conjunction with even and odd-harmonic synthesis, this attribute results in a preferred performance in subjective tests (with music and cinematic content) and with established perceptual metrics. We also present a technique for normalizing the input signal gain to a reference level using loudness metadata and level normalization between the LFE and L+R side chains (for cinematic content). A playback level compensation scheme is incorporated to maintain a proper low- and high-frequency balance at low and high-volume listening conditions (compared to a reference mid-volume condition). Future directions include the development of an adaptive shaping filter, $F(\omega)$, optimized for each content, and exploring a fusion metric based on combining *Rnonlin* and PEMO.

## 5 Acknowledgement

Samsung Electronics and Samsung Research America supported this work. The authors would like to thank Samsung's US Audio Lab staff, who helped with all aspects of this research, offered insightful suggestions, and contributed to this work.

## References

[1] Schouten, J., Ritsma, R., and Cardozo, L., "Pitch of the residue," *J. Acoust. Soc. Amer.*, 34(8), pp. 1418–1424, 1962.

[2] Zwicker, E. and Fastl, H., "Psychoacoustics: Facts and Models," *Springer (New York)*, 1999.

[3] Oxenham, A., "Revisiting place and temporal theories of pitch," *Acoust. Sci. Tech.*, 34(6), pp. 388–396, 2013.

[4] Larsen, E. and Aarts, R., "Audio Bandwidth Extension," *Wiley (New York)*, 2004.

[5] Lee, T., Lee, S., Park, Y.-C., and Youn, D. H., "Virtual bass system based on multiband harmonic generation," *Proc. 2013 IEEE Int. Conf. Consum. Elect. (ICCE)*, 2013.

[6] Shi, C., Mu, H., and Gan, W.-S., "A psychoacoustical preprocessing technique for virtual bass enhancement of the parametric loudspeaker," *Proc. 2013 IEEE Int. Conf. Acoust. Speech & Sig. Proc. (ICASSP)*, 2013.

[7] Oo, N., Gan, W.-S., and Lim, W.-T., "Generalized harmonic analysis of arc-tangent square root nonlinear device for virtual bass system," *Proc. 2013 IEEE Int. Conf. Acoust. Speech & Sig. Proc. (ICASSP)*, 2010.

[8] R. Giampiccolo, A. B. and Sarti, A., "A time-domain virtual bass enhancement circuital model for real-time music applications," *2022 IEEE 23$^{th}$ IEEE Wkshp. Mult. Sig. Proc. (MMSP)*, 2022.

[9] S. Cecchi, E. M. and Piazza, F., "A new approach to bass enhancement based on Prony's method," *2007 15$^{th}$ IEEE Int. Conf. Dig. Sig. Proc.*, 2007.

[10] P. Hoffmann, T. S. and Kostek, B., "Smart virtual bass synthesis algorithm based on music genre classification," *2014 IEEE Conf. Sig. Proc. (SPA 2014)*, 2014.

[11] S. Bharitkar, T., "Bandwidth extension with auditory filters and block-adaptive analysis," *Proc. 144$^{th}$ Audio Eng. Soc. Conv.*, 2018.

[12] N. Oo, W.-S. G. and Hawksford, M., "Perceptually-motivated objective grading of nonlinear processing in virtual-bass systems," *J. Audio Eng. Soc.*, 59((11)), pp. 804–824, 2011.

[13] C-T. Tan, N. Z., B. C. J. Moore and Mattila, V. V., "Predicting the perceived quality of nonlinearly distorted music and speech signals," *J. Audio Eng. Soc.*, 52((7/8)), 2004.

[14] Bai, M.-R. and Lin, W. C., "Synthesis and implementation of virtual bass system with a phase-vocoder approach," *J. Audio Eng. Soc.*, 11(54), pp. 1077–1091, 2006.

[15] Hill, A. J. and Hawksford, M. O. J., "A hybrid virtual bass system for optimized steady-state and transient performance," *2$^{nd}$ IEEE CEEC*, 2010.

[16] Zhang, S., Xie, L., Fu, Z.-H., and Yuan, Y., "A hybrid virtual bass system with improved phase vocoder and high efficiency," *9$^{t}$h IEEE ICSLP*, 2014.

[17] Mu, H., Gan, W.-S., and Tan, E.-L., "A psychoacoustic bass enhancement system with improved transient and steady-state performance," *Proc. 2012 IEEE Int. Conf. Acoust. Speech & Sig. Proc. (ICASSP)*, 2012.

[18] H. Mu, E.-L. T., W-S. Gan, "An objective analysis method for perceptual quality of virtual bass system," *IEEE/ACM Trans. Audio, Speech, Lang. Proc.*, 23(5), pp. 840–850, 2015.

[19] ITU-R Rec. BS. 1387-1, "Method for objective measurements of perceived audio quality," Standard, International Telecommunication Union, Geneva, CH, 2001.

[20] ISO 226:2003, "Acoustics — Normal equal-loudness-level contours," Standard, International Organization for Standardization, Geneva, CH, 2003.

[21] Huber, R. and Kollmeier, B., "PEMO-Q: A new method for objective audio quality assessment using a model of auditory perception," *IEEE/ACM Trans. Audio, Speech, Lang. Proc.*, 14(6), pp. 1902–1911, 2006.

[22] ITU-R Rec. BS. 775-4, "Multichannel stereophonic sound system with and without accompanying picture," Standard, International Telecommunication Union, Geneva, CH, 2022.

[23] I. Goodfellow, Y. B. and Courville, A., *Deep Learning*, 2016.

[24] J. Snoek, H. L. and Adams, R., "Practical Bayesian optimization of machine learning algorithms," *Proc. Neural Inf. Proc. Syst. (NIPS)*, 2012.

[25] Holman, T. and Kampmann, F., "Loudness Compensation: Use and Abuse," *J. Audio Eng. Soc.*, 26(7/8), pp. 804–824, 1978.