Audio Engineering Society

Convention Express Paper 438

Presented at the 160th Convention

2026 May 28-30, Copenhagen, Denmark

This Express Paper was selected on the basis of a submitted synopsis that has been peer-reviewed by at least two qualified anonymous reviewers. The complete manuscript was not peer reviewed. This Express Paper has been reproduced from the author's advance manuscript without editing, corrections or consideration by the Review Board. The AES takes no responsibility for the contents. 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.

Personalized Timbre Optimization for Stereophonic Sound Reproduction via Earphones: Part 2 – Practical Implementation and Validation on Consumer TWS Devices

Kimio Hamasaki[1, 2], Atsushi Hara [1], Haruto Hirai [1], Nao Tojo [1], Shun Saito [1], and Mitsuru Hosoo [1]

[1] *final Inc.*

[2] *ARTSRIDGE LLC*

Correspondence should be addressed to Kimio Hamasaki (hamasakik@final-inc.com)

ABSTRACT

This paper presents Part 2 of our study on personalized timbre optimization for stereophonic sound reproduction via earphones, following our previous work [1]. Part 1 established an auditory-model-based framework, hereafter referred to as PTTC-1 (Personalized Timbre Target response Curve), for reproducing the listener's individual reference for natural timbre. However, its practical application was constrained by the need for high-resolution 3D anatomical scanning, individualized HRTF computation by the boundary element method, and controlled listening tests conducted by trained audio engineers. The present paper introduces PTTC-2, a simplified and fully automated implementation for consumer True Wireless Stereo (TWS) earphones, in which a single smartphone image captures the listener's head and outer ear, the directionally averaged HRTFs are estimated by machine learning, and a brief in-app listening test determines a listener-specific weighting coefficient k.

Two empirical studies are reported. First, the analysis of k values selected by 2,850 users, together with a follow-up assessment of 29 listeners, indicates that k is a listener-specific parameter that is not strongly biased by program material. Second, a controlled subjective experiment with 24 listeners compared PTTC-1, PTTC-2, and two non-individualized target response curves used in consumer products. PTTC-1 and PTTC-2 did not differ significantly in holistic timbre attributes, and both were preferred over non-individualized target response curves. Principal component analysis further revealed that listener preference is governed by an overall naturalness axis rather than by a high-frequency emphasis axis. These results demonstrate that auditory-model-based timbre personalization can be effectively translated into a practical, consumer-ready technology that preserves the perceptual benefits of the full workflow.

1 Introduction

This paper presents Part 2 of our study on personalized timbre optimization for stereophonic sound reproduction via earphones, following our previous paper [1] presented at the AES International Conference on Headphone Technology in 2025. In Part 1, we proposed a novel auditory-model-based personalization framework designed to reproduce a listener's individual reference for natural timbre when listening to stereophonic content via earphones. Unlike conventional personalization approaches that rely on listener preferences or equalization heuristics, the proposed method isolated perceptually relevant timbral cues from spatial cues and optimized them using a perceptually grounded model of an individual auditory reference, fitted via controlled listening tests.

The subjective evaluation reported in Part 1 showed statistically significant improvements ($p < 0.01$) in perceived naturalness, richness, and clarity compared with earphones tuned to conventional non-individualized target response curves.

While Part 1 established the theoretical framework and perceptual validity of personalized timbre optimization—hereafter referred to as PTTC-1 (Personalized Timbre Target response Curve in Part 1)—the complexity of the required processing constrained its practical application. This paper, Part 2, addresses this constraint by introducing a simplified, fully automated implementation, hereafter referred to as PTTC-2, suitable for real-world deployment on consumer True Wireless Stereo (TWS) earphones.

Most research and commercial solutions for personalizing headphones and earphones have largely focused on optimizing spatial attributes such as localization, externalization, and envelopment, particularly for immersive binaural sound. In contrast, natural timbre reproduction—particularly in the context of conventional stereophonic playback, which still accounts for the majority of music consumed via earphones and headphones—has received considerably less dedicated attention [2, 3]. Inaccurate timbral fidelity can undermine the perceived musicality and emotional impact of music even when spatial attributes are well rendered, motivating dedicated study of timbre reproduction over earphones [2-4]. A substantial body of research has investigated listener preferences for sound quality and timbre, leading to the development of listener-preferred target response curves for headphones and in-ear headphones [5-7]. These curves represent general, non-individualized reference responses that improve average perceived audio quality. Traditional timbre personalization approaches, however, typically rely on equalization adjustments based on user preferences and do not account for individual auditory physiology or the perceptual mechanisms that underlie natural timbre perception. Theile [8, 9] characterized the behavior of the auditory system under natural listening conditions and proposed a perceptual distinction between spatial and timbral cues, in particular the Gestalt perception of timbre—a framework on which the present work builds.

Recent research has further refined target response curves for earphones. Knowles introduced a Preferred Listening Response Curve specifically addressing high-resolution audio playback above 10 kHz, a region the original Harman in-ear target did not adequately characterize [10]. Even when such target curves are carefully chosen, however, individual variability in ear-canal acoustics and pinna geometry produces residual deviations that cannot be removed by any single non-individualized response, motivating personalization beyond preference targets.

In parallel, research on individualized HRTF estimation has progressed rapidly through the use of machine learning. Image-based approaches, in which a neural network estimates the HRTF directly from photographs of the listener's pinna, have been shown to approximate measured HRTFs at substantially lower cost than acoustic measurements or boundary-element-method simulations [11, 12]. A recent comprehensive review by Bruschi et al. surveys the field and identifies trade-offs among acoustic, anthropometric, image-based, and mesh-based personalization strategies [13]. The PTTC-2 method proposed in this paper aligns with this image-based machine-learning branch but differs from prior work in two respects: (i) the goal is not full three-dimensional localization-grade HRTF reconstruction but rather the directionally averaged response that is required for natural-timbre personalization; and (ii) the method is designed for large-scale deployment on consumer TWS devices via a single smartphone capture.

The remainder of this paper is organized as follows. Section 2 reviews the auditory model and the original personalization method (PTTC-1) introduced in Part 1, providing the theoretical and procedural foundation for this work. Section 3 describes the simplified PTTC-2 implementation, including image-based HRTF estimation and the analysis of listener-specific weighting coefficients. Section 4 reports a controlled subjective listening experiment that compares PTTC-2 against PTTC-1, and two non-individualized target response curves. Section 5 discusses the implications of the findings, and Section 6 concludes.

2 Auditory model and personalization method (PTTC-1)

This section briefly summarizes the auditory model and the personalization method (PTTC-1) introduced in Part 1 [1], which together provide the theoretical and procedural foundation for the practical implementation (PTTC-2) presented in this paper. Whereas previous models for headphone and earphone reproduction have primarily addressed spatial attributes such as localization and

externalization, our model targets the spectral cues that govern timbre fidelity.

The model combines Theile's hearing framework [8, 9] with the insertion gain concept [15, 16] to personalize the target response curve at the Drum Reference Point (DRP). The transfer function $H(f)$ from a sound source to the Ear Reference Point (ERP)—defined as the plane at the entrance of the ear canal—is decomposed into a directional component $P(f)$ and a timbral component $T(f)$:

$$H(f) = P(f) \cdot T(f) \quad (1)$$

In earphone listening, the directional cues in $P(f)$ are bypassed by the absence of outer-ear spatial encoding, which degrades natural timbre perception. We therefore neutralize the directional component by setting $P(f) = 1$ and reconstruct the timbral component $T(f)$ from a directional power-spectrum average of $H(f)$ over a uniformly distributed grid of $N = 864$ measurement points (5.0° horizontal × 10.0° vertical), which provides sufficient resolution to capture the spectral variations of typical HRTFs:

$$H(f) = 1 \cdot T(f) = k \cdot \frac{1}{N} \sum_{i=1}^{N} HRTF_i(f) \quad (2)$$

To account for individual differences in auditory perception and ear anatomy, a listener-specific weighting coefficient k is introduced. Unlike preference-based equalization, k is an objective scaling factor obtained from controlled listening tests targeting the listener's reference for natural timbre, consistent with prior findings on listener variability in preferred frequency responses [5–7].

The Personalized Timbre Target response Curve (PTTC) at the DRP for the left-ear earphone is defined as:

$$PTTC^L(f) = k \cdot HRTF_{av}^L(f) + n \cdot TC_{low}(f) + m \cdot C^L(f) \quad (3)$$

where $HRTF_{av}$ is the listener-specific averaged HRTF in the mid-to-high frequency range (800 Hz – 10 kHz), TC_{low} is a non-listener-specific predefined target curve below 800 Hz, and C is a listener-specific ear-canal compensation curve. The 800 Hz division reflects the psychoacoustic finding that below this frequency, spatial cues exhibit minimal directional dependency, whereas above it, the spectrum strongly governs individual timbre perception. $PTTC^R$ is defined as in equation 3, where k , n , m , and TC_{low} are the same as left earphones, but $HRTF_{av}^R$ and C^R are specific to right earphones. As the contribution of C^L and C^R to perceived timbre is comparatively minor, it is not detailed in this paper. The original implementation of this model—hereafter referred to as PTTC-1—comprises four stages summarized in Figure 1.

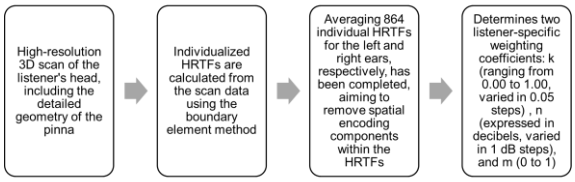


Figure 1. Flowchart for personalizing the target response curve of earphones to achieve natural timbre perception in stereophonic sound reproduction (PTTC-1).

Stages 1 and 2 acquire a high-resolution 3D scan of the listener's head, including detailed pinna geometry, and compute individualized HRTFs from the scan data by the boundary element method (BEM) over the 864 directions specified above.

Stage 3 averages the 864 directional HRTFs for each ear, yielding $HRTF_{av}^L$ and $HRTF_{av}^R$, which retain the spectral information required for natural timbre perception while eliminating directional encoding.

Stage 4 determines three listener-specific coefficients through controlled listening tests: k (0.00–1.00 in 0.05 steps), n (in 1 dB steps), and m (0–1). Figure 2 illustrates the effect of varying k on $k \cdot HRTF_{av}^L$ for Subject 09.

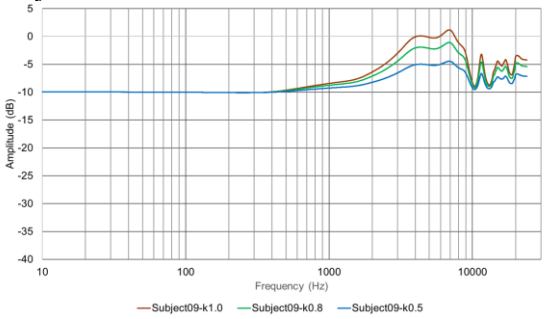


Figure 2. Example of $k \cdot HRTF_{av}^L$ of Subject 09 with different k values (1.0, 0.8, and 0.5).

The listening tests use four content categories selected to represent diverse timbral characteristics: (i) dialogue and a cappella vocals, (ii) solo musical instruments, (iii) a four-piece rhythm band without vocals, and (iv) vocals accompanied by a four-piece rhythm band. Figure 3 shows the resulting personalized target response curves for the left earphone of four representative subjects.

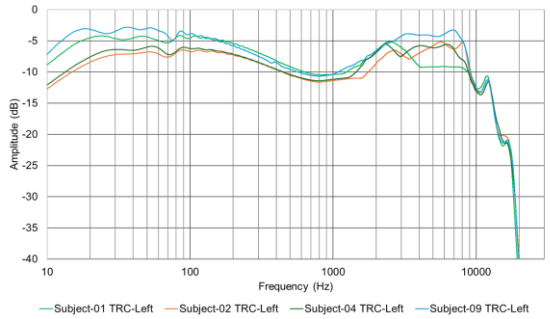


Figure 3. Personalized target response curve for the left earphone of four representative subjects.

3 Practical use of a personalized method for natural timbre (PTTC-2)

This section describes the practical implementation of the personalized natural-timbre method (hereafter referred to as PTTC-2), intended for deployment on consumer TWS earphones. As illustrated in Figure 4, PTTC-2 replaces the high-resolution 3D scanning step of PTTC-1 with a single capture of the listener's head and outer ear using a consumer smartphone, together with a reference sticker for dimensional calibration. The listener-specific HRTF parameters required for the auditory model in Section 2 are then estimated from these images by machine learning.

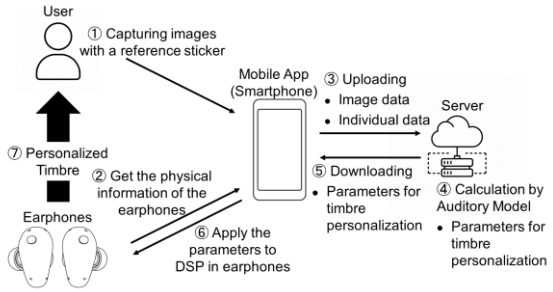


Figure 4. Workflow of the practical personalization method (PTTC-2): smartphone-based image capture of the listener's head and outer ear, with a reference sticker for dimensional calibration, followed by machine-learning-based estimation of HRTF parameters.

3.1 Estimation of $HRTF_{av}^{L}$ and $HRTF_{av}^{R}$

$HRTF_{av}^{L}$ and $HRTF_{av}^{R}$ are estimated from images of the listener's outer ear using machine learning, as shown in Figure 5. As illustrated in Figures 5 and 6, the estimation proceeds in two stages: nine anatomical feature points are first identified from the ear image, and from these feature points, the network predicts the frequency and amplitude of two spectral peaks (P1, P2) and one notch (N1) that characterize the listener's averaged HRTF. The $HRTF_{av}^{L}$ and $HRTF_{av}^{R}$ are then reconstructed from these estimated peak and notch parameters.

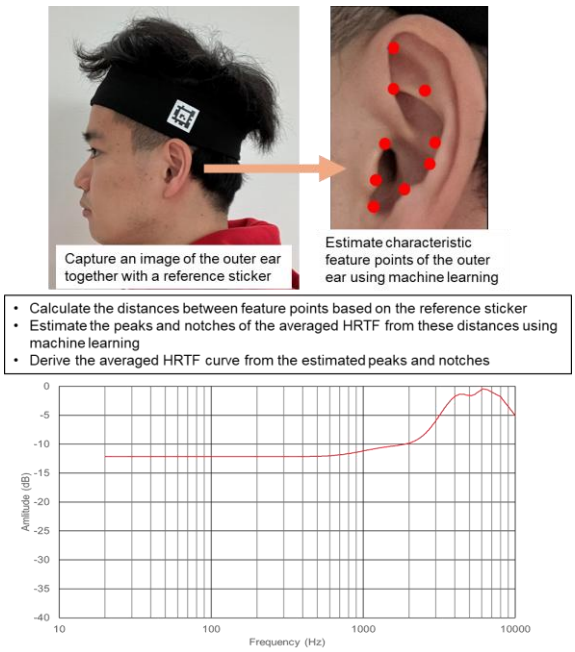


Figure 5. $HRTF_{av}^{L}$ estimated from an image of the outer left ear using machine learning.

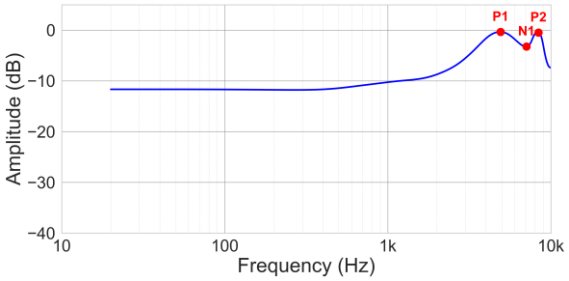


Figure 6. Two spectral peaks (P1, P2) and one notch (N1) are used to reconstruct the averaged HRTF.

Figure 7 compares the distributions of P1, N1, and P2 frequency and amplitude between PTTC-1 and PTTC-2. The distributions are narrower in PTTC-2 than in PTTC-1, indicating that the machine-learning-based estimation produces a tighter spread of estimated parameters than the BEM-based reference; nevertheless, the averaged HRTF obtained from PTTC-2 remains sufficiently accurate for natural-timbre personalization.

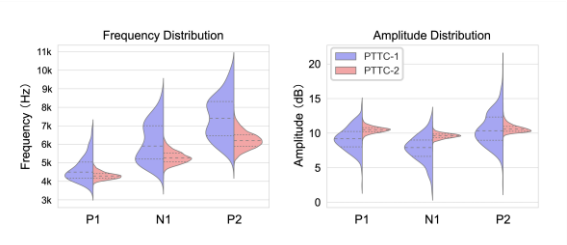


Figure 7. Comparison of the distributions of frequency and amplitude of P1, N1, and P2 between PTTC-1 and PTTC-2.

3.2 Listener-specific weighting coefficient k

Users of TWS earphones equipped with PTTC-2 perform a brief in-app listening test, in which they listen to their own preferred content and select the listener-specific weighting coefficient k that best matches their reference for natural timbre. Figure 8 shows the cumulative number of selections for each k value, based on 2,850 participants in this in-app listening test. The mean and median of the selected k values are 0.41 and 0.35, respectively. The fact that the mean exceeds the median indicates a positively skewed distribution, with most listeners selecting moderate-to-low values of k and a smaller proportion selecting high values.

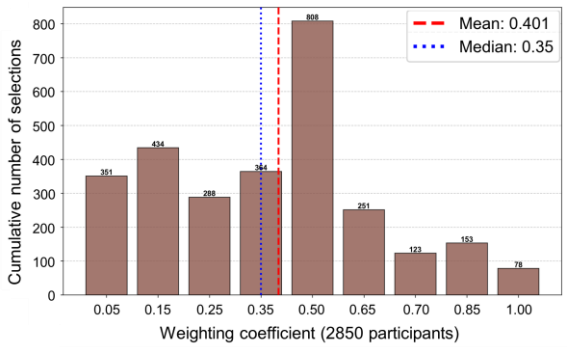


Figure 8. Cumulative number of selections for each value of the weighting coefficient k by 2,850 participants in the in-app listening test.

Because each user selected k based on personally preferred content, a follow-up subjective assessment was conducted to determine whether the acoustic features of the program material influenced their selection of k.

Twenty-nine participants took part in this assessment: 22 music students aged 19 to 26 and 7 audio researchers. All participants provided informed consent prior to the experiment, and the experimental protocol complied with the institutional ethical guidelines for research involving human subjects. Participants determined their preferred k value for each of three musical pieces (denoted A, B, and C) representing diverse genres. A is a pop song with vocals and a four-piece rhythm section; B is a violin sonata with piano accompaniment; and C is an electronic music piece with vocals. The acoustic features of A, B, and C are shown in Figure 9 and Table 1, and the resulting k values are reported in Table 2.

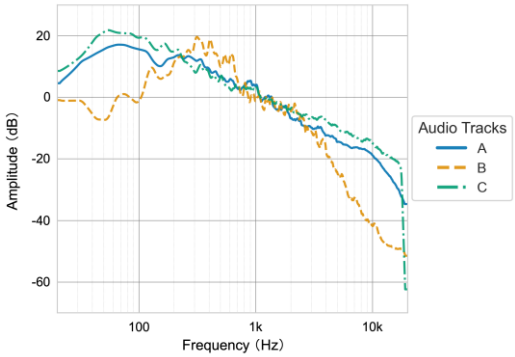


Figure 9. Long-Term Average Spectrum (LTAS) of the three musical pieces (A, B, and C) used in the subjective assessment.

Music	Spectral Centroid (Hz)	Low Energy Ratio (%)	Mid Energy Ratio (%)	High Energy Ratio (%)
A	2396.2	49.73	48.49	1.78
B	1051.78	6.76	93.19	0.06
C	2964.73	73.75	23.49	2.76

Table 1. Spectral features of musical pieces A, B, and C were used in the subjective assessment.

Music	Mean of weighting coefficient k	Range of weighting coefficient k
A	0.378	0.35 - 0.40
B	0.371	0.35 - 0.40
C	0.403	0.40 - 0.45

Table 2. Weighting coefficient k selected by each of the 29 participants for musical pieces A, B, and C.

Table 3 reports Pearson correlation coefficients between the acoustic features of pieces A, B, and C and the corresponding selected values of k. For all features, the absolute values of the correlation coefficients (|r|) are close to zero, and the corresponding p-values exceed 0.05, indicating no statistically significant linear relationship between the acoustic features of the program material and the selected k value. Given the moderate sample size (n = 29), this result should be interpreted as the absence of a strong content-dependent bias rather than as definitive proof of complete content independence. These results suggest that the spectral balance and

energy distribution of the program material do not strongly bias the value of k selected by listeners, supporting the assumption that k can be regarded as a listener-specific rather than a content-specific parameter.

Audio feature	Pearson r	p value	Significance
Spectral Centroid (Hz)	0.052	0.6314	Not Significant
Low Energy Ratio (%)	0.054	0.6168	Not Significant
Mid Energy Ratio (%)	-0.054	0.6168	Not Significant
High Energy Ratio (%)	0.054	0.6173	Not Significant

Table 3. Pearson correlation coefficients between the acoustic features of musical pieces A, B, and C and the selected weighting coefficients k.

4 Validation of practical use of a personalized method for natural timbre (PTTC-2)

The objective of this validation is to determine, through a quantitative subjective evaluation, whether the original timbre personalization method (PTTC-1) can be successfully applied to the practical use case (PTTC-2) and to clarify how four target response curves with different design philosophies affect the listener's perception of timbre.

4.1 Method of validation

A subjective listening experiment using the Semantic Differential (SD) method [17] was conducted. Twenty-four university students specializing in acoustics and audio engineering with normal hearing participated in the experiment. All participants provided informed consent prior to the experiment, and the experimental protocol complied with the institutional ethical guidelines for research involving human subjects. Four target response curves were compared: PTTC-1 (the original personalization method introduced in Part 1 [1] and reviewed in Section 2); PTTC-2 (the practical personalization method described in Section 3); the Diffuse Field Target Curve (DFTC) [14], which shares the same low-frequency response (below 800 Hz) as PTTC-1; and the Listener Preferred Target Curve (LPTC) [6]. All four curves were rendered through the same TWS earphone hardware so that only the target equalization differed across conditions. Figure 10 shows an example of the four target response curves for one participant.

The auditory stimulus consisted of vocals accompanied by a four-piece rhythm section. Participants rated their impressions on a 7-point bipolar scale (1 to 7) for 36 pairs of evaluation attributes (Table 4). Each pair consisted of two opposing descriptors, such as "natural" and "unnatural"; for example, when evaluating naturalness, a score of 1 indicates "very unnatural", 4 indicates "neutral", and 7 indicates "very natural".

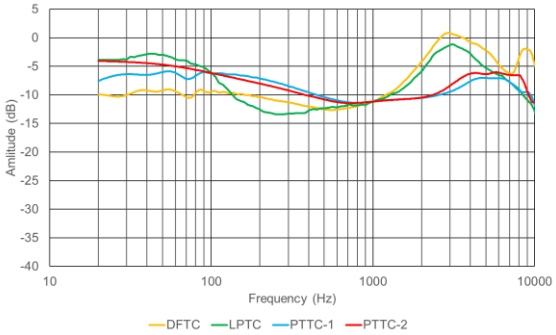


Figure 10. An example of the four target response curves (PTTC-1, PTTC-2, DFTC, LPTC) for one of the twenty-four participants.

Balanced	Pleasant	Noise-free	Diffused	Not dull	Bass+
Treble+	Not hissing	Not hard	Not hollow	Open	Clear
Not cracking	Bright	Not feeble	Soft	Clear	Natural
Presence	Clean	Space	Not Jarring	Not too loud	Gentle
Not dry	Thick	Not sharp	Wide band	Transparent	Smooth
Non-fatigue	Rich	Good timbre	Good sound	Enjoyable	Like

Table 4. List of the 36 bipolar evaluation attributes used in the SD method.

To uncover the perceptual structure underlying these ratings, principal component analysis (PCA) without rotation was applied; sampling adequacy was verified by the Kaiser–Meyer–Olkin (KMO) measure. To test whether the four target conditions yielded statistically distinguishable ratings, the non-parametric Kruskal–Wallis test was applied to each evaluation attribute. Where a significant main effect was observed, post-hoc pairwise comparisons were performed using the Dwass–Steel–Critchlow–Fligner (DSCF) procedure.

4.2 Overall results of the SD evaluation

Figure 11 shows the SD evaluation results, presenting mean scores and 95% confidence intervals for each of the four target response curves across the twenty-four subjects. Both PTTC-1 and PTTC-2 consistently obtained higher ratings than DFTC and LPTC for the attributes related to overall sound quality and listener preference, indicating that the personalized natural-timbre method can be effectively applied in practical use cases.

4.3 Perceptual structure revealed by PCA

The KMO value for the overall sample was 0.887, indicating that the dataset is well-suited to PCA. The first two principal components together accounted for 54.5% of the total variance (Table 5, Figure 12).

PC1 (42.1% of variance) was strongly loaded by overarching positive evaluation attributes such as "Good timbre" (0.882), "Good sound" (0.863), "Enjoyable" (0.853), "Like" (0.843), and "Natural" (0.827). The very high loadings on this set of attributes suggest that listeners formed an integrated, holistic impression of overall sound quality rather than evaluating individual attributes in isolation. The high loading for "Natural" further indicates that this notion of "Good sound" was closely identified with the absence of artificial spectral emphasis. The simultaneous high loadings of "Enjoyable" and "Like" with the timbral attributes show that affective engagement was tightly coupled with perceived sound quality, consistent with the heuristic that "good sound enables musical immersion".

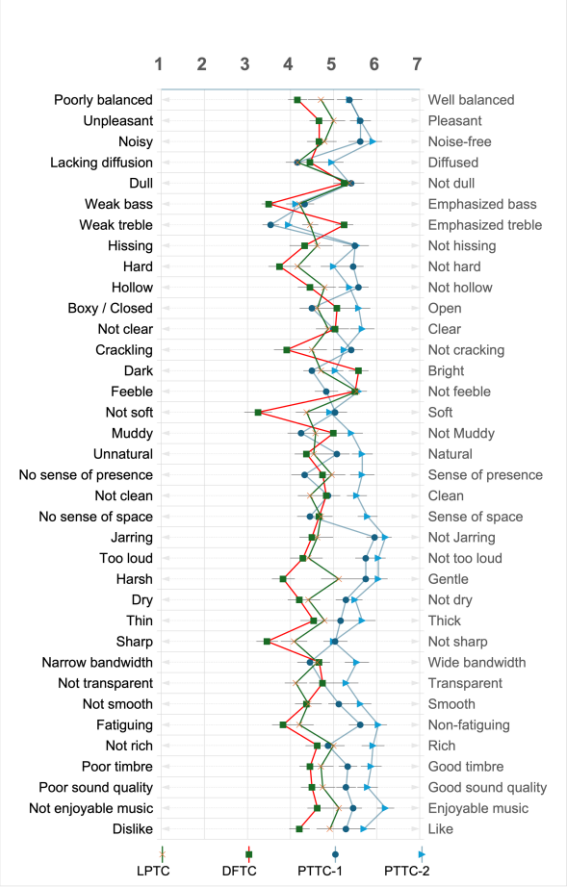


Figure 11. Mean subjective ratings and 95% confidence intervals of the 36 evaluation attributes for the four target response curves (LPTC, DFTC, PTTC-1, and PTTC-2).

PC2 (12.4% of variance) reflected the perceived high-frequency character. Its positive direction was driven by "Treble+" (0.726) and "Bright" (0.647), whereas its negative direction was driven by "Gentle" (−0.440), "Not sharp" (−0.442), and "Soft" (−0.403). PC2, therefore, represents a continuum from a bright, edgy character at the top to a soft, comfortable character at the bottom. While moderate high-frequency emphasis can enhance perceived clarity, its excess is also a primary source of perceived sibilance and harshness, placing PC2 on a perceptual trade-off between resolution and comfort.

Summary

Component	SS Loadings	% of Variance	Cumulative %
1	15.16	42.11	42.1
2	4.47	12.41	54.5

Table 5. Summary statistics of the principal components.

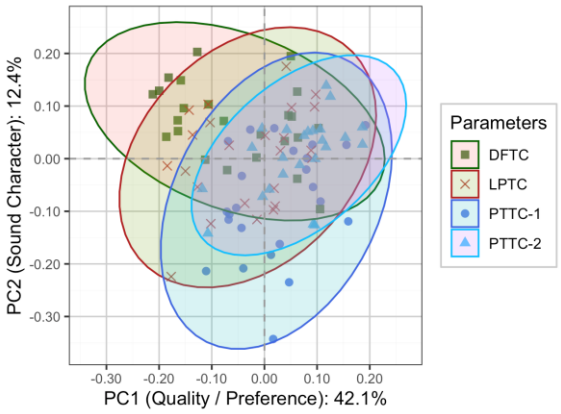


Figure 12. PCA loading plot of the 36 evaluation attributes across the four target response curves. PC1 explains 42.1%, and PC2 explains 12.4% of the total variance.

4.4 Statistical comparison among the four target curves

The Kruskal–Wallis test revealed statistically significant differences among the four target response curves ($p < .05$ or $p < .001$) in more than 30 of the 36 evaluation attributes. The largest effect sizes were observed for the three attributes, as shown in Table 6. These results indicate that perceived high-frequency strength and long-term listening comfort are the

attributes most strongly affected by the choice of target response curve.

Evaluation Term	X^2	ε^2	p-value
Treble+	34.474	0.36288	<.001 ***
Non-fatigue	31.253	0.32898	<.001 ***
Not too loud	23.291	0.24517	<.001 ***
	* p < .05, ** p < .01, *** p < .001		

Table 6. Three evaluation attributes with the largest effect sizes (χ^2 and ε^2) in the Kruskal–Wallis test among the four target response curves.

4.5 PTTC versus DFTC and LPTC

In the PCA loading plot of Figure 12, both PTTC-1 and PTTC-2 are located on the positive side of PC1 and form a high-rated group, whereas DFTC and LPTC are located on the negative side. The two PTTC conditions did not differ significantly from each other on any holistic-preference attribute. DSCF post-hoc comparisons (Table 7) showed that the PTTC group differed significantly from DFTC and LPTC on "Non-fatigue", "Not Jarring" ($p < .001$), and "Rich" ($p < .005$). These results indicate that the two PTTC variants confer both higher overall preference and superior long-term listening comfort.

Evaluation Term	Comparison	DSCF W	DSCF p-value
Non-fatigue	DFTC vs PTTC-1	4.91	0.003 **
	DFTC vs PTTC-2	5.87	<.001 ***
	LPTC vs PTTC-1	4.91	0.003 **
	LPTC vs PTTC-2	6.28	<.001 ***
Not Jarring	DFTC vs PTTC-1	3.952	0.027 *
	DFTC vs PTTC-2	4.653	0.006 **
	LPTC vs PTTC-1	4.469	0.009 **
	LPTC vs PTTC-2	5.283	0.001 ***
Rich	DFTC vs PTTC-2	4.673	0.005 **
	LPTC vs PTTC-2	3.755	0.04 *
	* p < .05, ** p < .01, *** p < .001		

Table 7. DSCF post-hoc comparisons of the PTTC (PTTC-1, PTTC-2) against DFTC and LPTC for representative comfort- and preference-related evaluation attributes.

4.6 Difference between PTTC-1 and PTTC-2

Although PTTC-1 and PTTC-2 were grouped together as the high-rated cluster, a closer examination revealed perceptual nuances that distinguish the two variants. Both showed very high loadings on the holistic timbre attributes of PC1, and no significant difference was observed for "Good timbre" ($p = .245$), "Natural" ($p = .478$), or "Smooth" ($p = .467$). The absence of audible distortion and the overall smoothness of the spectral character were therefore perceived as essentially equivalent between PTTC-1 and PTTC-2.

In contrast, statistically significant differences ($p < .05$) emerged on attributes related to spatial impression and timbral richness, such as "Rich" ($p = .036$), "Space" ($p = .015$), "Wide band" ($p = .044$), "Presence" ($p = .018$), and "Enjoyable" ($p = .027$). On all of these attributes, PTTC-2 received the highest mean rating. Listeners therefore treated PTTC-1 and PTTC-2 as equivalent with respect to the holistic notion of "Good timbre" but distinguished PTTC-2 by a denser timbral texture and a stronger sense of acoustic space, suggesting greater musical immersion.

4.7 Summary

The validation experiment supports two conclusions of practical relevance to the proposed method. First, perceived sound quality in earphone listening reduces to two interpretable principal axes: an overall preference axis (PC1) dominated by "naturalness", and a high-frequency character axis (PC2) corresponding to a brightness-versus-comfort trade-off. Second, a clear perceptual hierarchy emerges among the four target response curves: PTTC-1 ≈ PTTC-2 > LPTC > DFTC. The simplified PTTC-2 procedure is statistically indistinguishable from the full PTTC-1 method on overall preference and timbral naturalness, while exhibiting comparable or somewhat richer spatial and timbral character. This outcome confirms that the smartphone-based PTTC-2 implementation preserves the perceptual benefits of the full personalization workflow on consumer TWS earphones.

5 Discussion

The findings carry several implications. First, the large-scale data of 2,850 listeners reported in Section 3 show that the listener-specific weighting coefficient k spans a wide range, indicating substantial listener-to-listener variability that no single non-individualized target response curve can accommodate. The follow-up assessment with 29 listeners further showed no statistically significant correlation between the acoustic features of the program material and the selected k value, supporting the assumption that k can be regarded as a listener-specific rather than a content-specific parameter, and justifying the design choice of fitting k once per listener and reusing it across content.

Second, the validation experiment in Section 4 directly addresses the central question of whether the simplification from PTTC-1 to PTTC-2 preserves the

perceptual benefit of the original method. PTTC-1 and PTTC-2 did not differ significantly on holistic timbre attributes such as "Good timbre" ($p = .245$), "Natural" ($p = .478$), and "Smooth" ($p = .467$). The smartphone-based image capture and the machine-learning-based estimation of $HRTF_{av}^{L}$ and $HRTF_{av}^{R}$ thus preserve the dominant perceptual benefit of PTTC-1, despite producing a narrower distribution of estimated peak-and-notch parameters than the BEM-based reference (Figure 7). The narrower distribution suggests that the machine-learning estimator regresses toward the training-set mean; importantly, this regression does not impair perceived naturalness, indicating that the auditory model is robust to the residual estimation error of image-based HRTF estimation.

Third, the hierarchy PTTC-1 ≈ PTTC-2 > LPTC > DFTC admits a coherent perceptual interpretation through the two PCA axes. PC2 (high-frequency character) is decoupled from PC1 (overall naturalness and listener preference), and listener preference is governed by PC1 rather than by PC2. Strong high-frequency emphasis—often associated with apparent clarity—therefore does not by itself drive listener preference and may indeed harm it over long listening durations, as shown by the low ratings of DFTC on "Non-fatigue" and "Not Jarring". This finding aligns with recent work on in-ear target response curves [10], which similarly emphasizes perceptual balance over mere extension of the high-frequency response.

Several limitations should be acknowledged. The validation used 24 listeners, a single TWS hardware platform, and a single vocal program accompanied by a four-piece rhythm section. The content-independence of k was assessed with 29 listeners; the moderate sample size limits statistical power, and the result should be interpreted as the absence of a strong content-dependent bias rather than as proof of complete content independence. The machine-learning estimator for $HRTF_{av}^{L}$ and $HRTF_{av}^{R}$ was evaluated indirectly through the perceptual outcome of PTTC-2; a direct objective benchmark against measured HRTFs, together with broader evaluation across listeners, programs, and hardware, and extension to binaural and immersive content, are left to future work.

6 Conclusions

This paper presented PTTC-2, a practical implementation of the auditory-model-based timbre personalization method introduced in Part 1, suitable for deployment on consumer TWS earphones. PTTC-2 replaces the high-resolution 3D anatomical scanning, BEM-based HRTF computation, and trained-engineer listening tests of PTTC-1 with a single smartphone capture of the listener's head and outer ear, machine-learning-based estimation of the directionally averaged $HRTF_{av}^{L}$ and $HRTF_{av}^{R}$, and a brief in-app listening test in which the listener selects the listener-specific weighting coefficient k.

Three principal results were reported. First, the analysis of selected weighting coefficients from 2,850 users, together with a follow-up assessment by 29 listeners of three musical pieces, indicates that k is a listener-specific parameter that is not strongly influenced by the acoustic features of the program material. Second, the controlled subjective listening experiment with 24 listeners showed that PTTC-2 is statistically indistinguishable from PTTC-1 on holistic timbre attributes—including "Good timbre", "Natural", and "Smooth"—and that both PTTC-1 and PTTC-2 are preferred over the conventional non-individualized target response curves, establishing the perceptual hierarchy PTTC-1 ≈ PTTC-2 > LPTC (Listener Preferred Target Curve) > DFTC (Diffuse Field Target Curve). Third, principal component analysis revealed that listener preference is governed by an overall naturalness axis (PC1) rather than by a high-frequency emphasis axis (PC2), explaining why DFTC, despite providing the strongest perceived high-frequency content, was rated lowest on long-term listening comfort.

Together, these results demonstrate that the perceptual benefits of the auditory-model-based personalization framework introduced in Part 1 can be retained in a workflow simple enough for large-scale deployment on consumer hardware, and they support the broader conclusion that the central determinant of preferred timbre on earphones is the listener-specific naturalness of the spectral envelope rather than its high-frequency emphasis. Extension of the framework to other program materials and hardware platforms, direct objective benchmarking of the machine-learning HRTF estimator, and integration with personalized spatial rendering for binaural and immersive content remain promising directions for future work.

7 Acknowledgments

The authors thank the participants of the listening experiments for their valuable contributions. We also thank our colleagues at final Inc. for their support of this work.

References

[1] K. Hamasaki, N. Tojo, A. Hara, H. Hirai, S. Saito, and M. Hosoo, "Personalized Timbre Optimization Based on a New Auditory Model for Stereophonic Sound Reproduction via Earphones," *AES International Conference on Headphone Technology*, Helsinki, Finland, 2025.

[2] T. Lokki, V. Pulkki, M. Karjalainen, and L. Savioja, "Creating audibly convincing virtual acoustics and 3D audio: Past, present, and future," *IEEE Signal Process. Mag.*, vol. 28, no. 6, pp. 18–28, 2011.

[3] F. Rumsey, *Spatial Audio*, Focal Press, 2001.

[4] B. C. J. Moore and B. R. Glasberg, "Modeling binaural loudness," *J. Acoust. Soc. Am.*, vol. 121, no. 3, pp. 1604–1620, 2007.

[5] S. E. Olive, T. Welti, and E. McMullin, "Listener Preference for In-Room Loudspeaker and Headphone Target Responses," presented at the *AES Convention*, New York, USA, Paper 8994, 2013.

[6] S. E. Olive and T. Welti, "Factors that Influence Listeners' Preferred Bass and Treble Balance in Headphones," presented at the *AES Convention,* New York, USA, Paper 9382, 2015.

[7] S. E. Olive, T. Welti, and O. Khonsaripour, "The Preferred Low Frequency Response of In-Ear Headphones," presented at the *AES Conference on Headphone Technology*, Aalborg, Denmark, 2016.

[8] G. Theile, "Equalization of studio monitor headphones," presented at the *AES Conf. on Headphone Technology*, Aalborg, Denmark, 2016.

[9] G. Theile, "On the Standardization of the Frequency Response of High-Quality Studio Headphones," *J. Audio Eng. Soc.*, vol. 34, no. 12, pp. 956–969, 1986.

[10] Knowles Corporation, "A Preferred Listening Response Curve for High Resolution In-Ear Headphones," *Technical White Paper*, 2022.

[11] N. Javeri, P. B. Dutta, K. Sunder, and K. Jain, "Predicting Personalized Head-Related Transfer Functions Using Acoustic Scattering Neural Networks," presented at the *AES 153rd Convention*, 2022.

[12] R. Miccini and S. Spagnol, "HRTF Individualization using Deep Learning," in *Proc. IEEE Conference on Virtual Reality and 3D User Interfaces Workshops (VRW)*, Atlanta, USA, 2020, pp. 390–393.

[13] V. Bruschi, L. Grossi, N. A. Dourou, A. Quattrini, A. Vancheri, T. Leidi, and S. Cecchi, "A Review on Head-Related Transfer Function Generation for Spatial Audio," *Applied Sciences*, vol. 14, no. 23, p. 11242, 2024.

[14] *ITU-R BS.708*, "Determination of the electro-acoustical properties of studio monitor headphones," 1990.

[15] C.J. Struck, S. Temme, "Headphone Response: Target Equalization Trade-offs and Limitations," *AES convention* paper 9374, New York, USA, 2015.

[16] C.J. Struck, "Free Plus Diffuse Sound Field Target Earphone Response Derived From Classical Room Acoustics Theory," *AES convention* paper 8993, New York, USA, 2013.

[17] C.E. Osgood, G.J. Suci, P.H. Tannenbaum, *The measurement of meaning*, University of Illinois Press, Urbana, 1957.