Perceptually Optimized Personal Sound Zones

EE 395 Final Project

Tanay Mannikar

March 14, 2025

1 Introduction

Acoustic equalization is one of the most well-known applications of adaptive filtering. Common equalization tasks include room transfer function (RTF) compensation, which aims to enhance or mitigate the room's reverberant response or sound field in the frequency domain. Here, we examine a special case of this paradigm known as personal sound zones (PSZs), a form of localized sound field control or dereverberation. Personal sound zones aim to isolate sound within a region of space while mitigating the presence of that sound present in one or more targeted nearby regions, respectively known as bright and dark zones. By creating these effective acoustic bubbles, one can design personal spaces in which the system simultaneously delivers an audio source or program to one zone while other nearby sound zones are also active, allowing isolation of audio content and information streaming without the need for headphones. These spaces are of immense interest in applications such as automotive cabins, where drivers and passengers seated in close proximity are able to listen to and view individualized content without disturbing one another. In this way, important alerts can also be directed to only the driver or individual passengers, or speech programs can be streamed in multiple different languages for improved intelligibility, among other applications. However, the complex geometries of real-life listening spaces and vehicle cabins pose a complex physical optimization problem.

The generation of static PSZs is practically implemented using adaptive signal processing algorithms that play a critical role in defining optimization criteria for the localized manipulation of sound pressure fields and in facilitating their real-time implementation. In recent times, focus has also shifted towards nonlinear adaptive filters for moving head-tracked sound zones in the form of neural networks [1], as deep neural network filter operations can now be deployed on embedded hardware and software within real-time constraints. Furthermore, optimization schemes have started to include perceptual criteria into their objectives, allowing PSZ filters to be computed with respect to direct input from the audio programs. Perceptual constraints account for psychoacoustic phenomena such as masking and open the door for enhancing or even involving subjective criteria such as speech and audio quality metrics. These methods allow systems to emphasize relevant time-frequency information and physical parameters that address the understanding of auditory perception, increasing both efficiency and subjective listening performance in real-world applications. In this project, the goal is to outline important algorithms and practical applications of perception-oriented adaptive filtering in personal sound field control.

2 A Perceptual Primer

2.1 Psychoacoustic Masking

Psychoacoustic or auditory masking is the backbone of modern perceptual audio coding and compression algorithms such as MPEG-3. Frequency masking refers to the absence of perceivable sinusoidal components within a signal due to nearby louder tones. More generally, masking in both time and frequency can occur due to competing spectrotemporal information altering the perception of sounds existing on their own. In the frequency domain, signals can be broken down by one or more parallel filterbanks which split the audio into bandpassed components spanning the audible frequency range. This allows the ability to reduce the information, i.e. number of bits, required to represent frequency components surrounding a louder tone within the bandwidth of an auditory filter without suffering from quantization error. In perceptual audio compression, this effect is demonstrated by the *13 dB miracle* - a noisy audio signal with a signal-to-noise ratio (SNR) of 13 dB is perceived with clearly audible artifacts and distortion when broadband white noise is applied [2]. However, shaping the noise by auditory filtering and placing the noise carefully into masked regions of the signal almost completely mitigates the perceived distortion even though the SNR remains identical to standard white noise. Thus, it is apparent that in the context of personal sound zones where audio information is being streamed to target points in space, standard signal power distortion metrics such as SNR alone are not sufficient to represent the subjective performance of the system. As will be discussed in Section 3.1, adaptively applying inverted masking curves computed frame-by-frame from a target audio signal to weight the error of the signal power generated within a sound zone during optimization is a valuable implementation of this concept.

2.2 Missing-Fundamental Phenomenon

The missing-fundamental phenomenon arises in the concept of pitch perception. Humans perceive pitch not only at a single frequency, but also as a set of integer multiples, or overtones, of the loudest frequency component associated with any given pitch. Conventionally, this fundamental frequency f_0 is set to the lowest component so that only overtones are considered separate, but lower factors can also be considered - these are known as the *sub-harmonics* of the fundamental frequency. *Virtual pitch* is then defined as the harmonic series of f_0 with f_0 absent, and is nearly equivalent in perceptual quality to a pitch which includes the fundamental. This quality is known as the *missing-fundamental phenomenon* and can be exploited to minimize the importance of f_0 and its harmonics in practical scenarios. In short, any given pitch can be characterized as the set of frequencies

$$A = \{nf_0 \mid n > 0\}, \quad 20\text{Hz} \leq f_0 < nf_0 \leq 20\text{kHz}$$

while virtual pitch is defined as

$$\hat{A} = \{nf_0 \mid n > 1\}, \quad 20\text{Hz} \leq f_0 < nf_0 \leq 20\text{kHz}$$

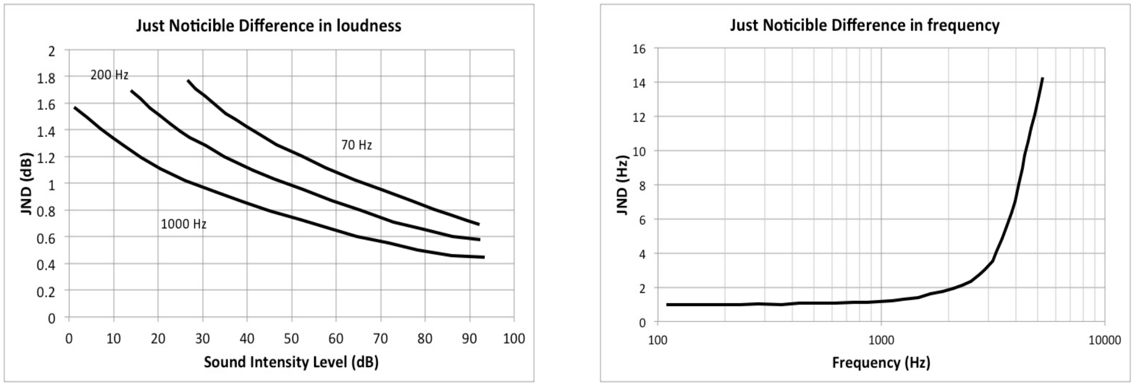
where $\hat{A} \simeq A$ for most pitches present in speech and music within the audible frequency range above a certain threshold. In fact, harmonics are commonly amplified in practice to improve perception of low frequency components.

In PSZ setups, loudspeaker arrays are the primary apparatus used to generate bright and dark zones. In the case of line arrays, multichannel audio output from adjacent loudspeakers contains overlapping frequency content adjusted by some time-varying gain factor in order to produce the desired pressure field in dark and bright zones. However, this geometry leads to a comb filtering effect at positions not directly centered in front of the line array (assumed symmetrical about the midline) due to the feedforward effect of time delays associated with multiple copies of the same signal reaching the same location causing constructive and destructive interference, leading to perceived distortion known as "flanging" in digital audio. An adaptive filtering solution to this problem is discussed in Section 3.2. While PSZ implementations normally assign a discretized grid of target points within a region of space, points corresponding to the location of either ear can undergo further optimization in a similar manner by taking into account binaural perception, or the effects of phase and group delay of the resulting signal arriving at both ear locations, which is detailed in the next section.

2.3 Just-Noticeable Difference (JND)

We now turn to thresholds of sound intensity and frequency in both monaural and binaural cases. The Just-Noticeable Difference (JND) characterizes the minimum threshold at which humans perceive a change in stimulus. In the monaural case, JND-dB curves reveal the minimum change in dB at each frequency that is perceived as a change in stimulus. This information can be employed to introduce a relaxation in the constraint of magnitude error at each frequency bin for a PSZ filter, or by applying the mean of each curve across the frequency range. Similarly, JND-Hz curves identify the minimum difference required to discern a perceptual difference in frequency. Figures 1(a) and (b) taken from [3] provide example JND-dB and JND-Hz curves.

For binaural perception, JND of the interaural time difference (ITD) represents the minimum perceptible time delay for a sound arriving at both ears to be perceived as two separate signals, which varies between 30 and 200 μs. JND of the interaural loudness difference (ILD) provides the minimum perceptible shift in intensity at each



(a) JND-dB

(b) JND-Hz

Figure 1: Monaural JND thresholds

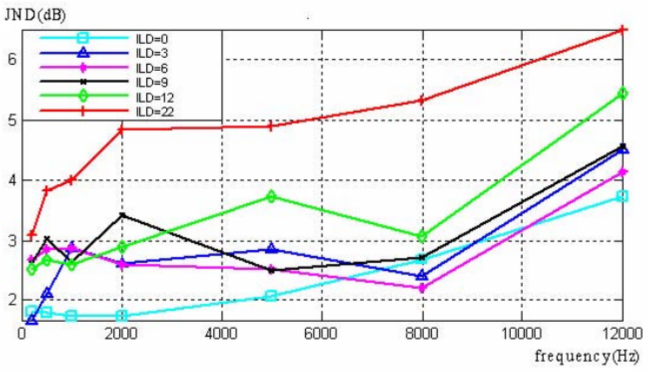


Figure 2: JND-ILD thresholds

frequency perceived between stimuli either ear. An example plot of experimental JND-ILD curves from [4] is provided in Figure 2. While the monaural JND is applicable to characterize the magnitude and phase distortion of isolation performance at a single zone point, the latter two provide binaural thresholds for error at points representing locations of either ear. The ability to vary group delay across the frequency range within a certain threshold in stereo signals without a perceived difference is applied in the filtering approach described in Section 3.2 to further improve time-varying multichannel decorrelation.

3 Adaptive Algorithms

First, the error objective for the generation of static PSZs without perceptual constraints is defined as given in [5]. To solve for the pressure $p_m[n]$ at control point m and time index n generated by loudspeaker l, the linear convolution of input signal $x[n]$ with room impulse responses (RIRs) $\boldsymbol{h}_{ml}$ associated with each loudspeaker position and control point gives the uncontrolled pressure

$$\boldsymbol{y}_{ml}[n] = \boldsymbol{X}[n]\boldsymbol{h}_{ml} \tag{1}$$

to which the $L \times J$ matrix $\boldsymbol{q}$ of FIR filters with length J is applied to the output of L loudspeakers to obtain the optimal pressure at point m

$$p_m[n] = \sum_{l=1}^{L} \boldsymbol{y}_{ml}^T[n]\boldsymbol{q}_l = \boldsymbol{y}_m^T[n]\boldsymbol{q} \tag{2}$$

where $\boldsymbol{X}[n]$ is constructed as outlined in [5]. The desired signal $d_m[n] = (h_{mz} * x)[n]$ in the set of bright zone control points $z \in \mathcal{M}_B$ and $d[n] = 0$ in the set of dark zone control points $z \in \mathcal{M}_D$ correspond to the pressure generated for a single zone with no interference in the bright zone case and no pressure in the dark zone case. The error is then defined as

$$\varepsilon_m[n] = d_m[n] - p_m[n]. \tag{3}$$

3.1 AP-VAST

The complete implementation for adaptive and perceptually optimized variable span trade-off (AP-VAST) filtering in full detail is outlined in [5], as visually depicted by Figure 3. By introducing an inverted psychoacoustic masking curve computed from the desired signal $d_m[n]$ (to avoid iterative procedures) as a weighting filter $w_m[n]$ to the error formulation, we modify the error from (3) to obtain

$$(\varepsilon_m * w_m)[n] = \tilde{\varepsilon}_m[n] = \tilde{d}_m[n] - \tilde{p}_m[n] \tag{4}$$

for the AP-VAST approach. This masking filter can be computed as outlined in [6]. Note that by fixing $w_m[n] = w[n]$ as a static filter obtained by averaging many masking curves obtained by real-world speech and audio signal segments, the error is identical to

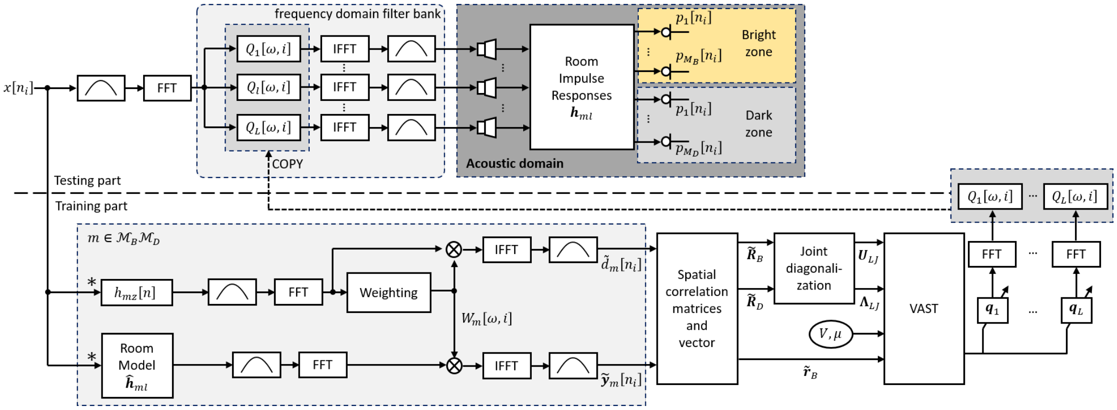


Figure 3: VAST architecture

that of P-VAST (i.e., static masking instead of adaptive masking) and is independent of control point location. The rest of the solution is equivalent to the standard VAST approach with these perceptual modifications. The signal distortion power (SDP) in the bright and dark zones is then defined as

$$\tilde{\mathcal{S}}_{\mathrm{B}}(\boldsymbol{q}) = \tilde{\sigma}_d^2 - 2\boldsymbol{q}^T\tilde{\boldsymbol{r}}_{\mathrm{B}} + \boldsymbol{q}^T\tilde{\boldsymbol{R}}_{\mathrm{B}}\boldsymbol{q}, \tag{5}$$

$$\tilde{\mathcal{S}}_{\mathrm{D}}(\boldsymbol{q}) = \boldsymbol{q}^T\tilde{\boldsymbol{R}}_{\mathrm{D}}\boldsymbol{q}, \tag{6}$$

where

$$\tilde{\sigma}_d^2 = \frac{1}{|\mathcal{M}_{\mathrm{B}}|N}\sum_{n=0}^{N-1}\sum_{m\in\mathcal{M}_{\mathrm{B}}}|\tilde{d}_m[n]|^2, \tag{7}$$

$$\tilde{\boldsymbol{r}}_{\mathrm{B}} = \frac{1}{|\mathcal{M}_{\mathrm{B}}|N}\sum_{n=0}^{N-1}\sum_{m\in\mathcal{M}_{\mathrm{B}}}\tilde{\boldsymbol{y}}_m[n]\tilde{d}_m[n], \tag{8}$$

$$\tilde{\boldsymbol{R}}_{\mathrm{C}} = \frac{1}{|\mathcal{M}_{\mathrm{C}}|N}\sum_{n=0}^{N-1}\sum_{m\in\mathcal{M}_{\mathrm{C}}}\tilde{\boldsymbol{y}}_m[n]\tilde{\boldsymbol{y}}_m^T[n], \tag{9}$$

for N samples. The VAST algorithm solves the optimal filtering problem $\boldsymbol{q}$ by generalized eigenvalue decomposition of the spatial correlation matrices $\boldsymbol{R}_C$ due to their real, positive semidefinite and symmetrical nature in the bright and dark zone control points B and D for the desired zone $C \in \{B, D\}$. By approximating $\boldsymbol{q} \simeq \boldsymbol{U}_V\boldsymbol{a}_V$ where $\boldsymbol{U}_V$ is the matrix of V generalized eigenvectors of the jointly diagonalized spatial correlation matrices and $\boldsymbol{a}$ is a V coefficient vector, solving the optimization problem for the Lagrangian

$$\mathcal{L}(\boldsymbol{q}) = \tilde{\mathcal{S}}_{\mathrm{B}}(\boldsymbol{q}) + \mu(\tilde{\mathcal{S}}_{\mathrm{D}}(\boldsymbol{q}) - \epsilon) \tag{10}$$

where μ is the Lagrange multiplier gives the optimal control filter

$$\boldsymbol{q}_o(V,\mu) = \boldsymbol{U}_V\boldsymbol{a}_o(V,\mu) = \sum_{v=1}^{V}\frac{\boldsymbol{u}_v^T\tilde{\boldsymbol{r}}_B}{\lambda_v+\mu}\boldsymbol{u}_v, \quad 1 \le V \le LJ \tag{11}$$

where λ_v and $\boldsymbol{u}_v$ are the vth eigenvalue and eigenvector. Parametrizing the choice of V

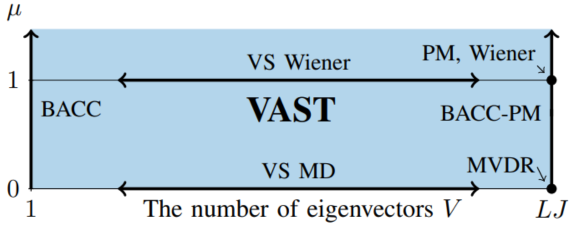


Figure 4: VAST parameter plane

and μ allows one to obtain many of the existing adaptive solutions, allowing the trade-off of advantages of either solution depending on the desired physical or perceptual evaluation metric. A visual representation of these relations from [5] is provided in Figure 4.

From the listening tests conducted in the study, an improvement by over 20% is achieved by AP-VAST compared to the standard VAST approach. Interestingly, the P-VAST implementation performs best in overall subjective testing, but the authors speculate that allowing variation of V and μ for consecutive audio segments might allow better performance for the AP-VAST system. Thus, while physical performance metrics of the AP-VAST system are not improved over VAST, the perceptual quality of the generated field is greatly improved, meaning perceptual considerations are paramount to the generation of ideal personal zound zones as opposed to solely targeting minimization of signal power and/or acoustic contrast (AC).

3.2 Adaptive Multichannel Decorrelation

As stated in Section 2.2, multichannel signal correlation, or the comb filtering effect as a result of the loudspeaker array geometry is addressed. Prior to PSZ filter generation, adaptive notch filtering to the audio signal can be applied in order to decorrelate the fundamental frequency and harmonics present in adjacent copies of the multichannel output as proposed in [7]. However, this approach assumes that a dominant harmonic structure exists in the program (i.e. a primary speaker in a speech mixture).

Common approaches to adaptively tracking the fundamental frequency f_0 involve autocorrelation to identify the most relevant lags as the time delays associated with the fundamental, instantaneous frequency estimates from the instantaneous phase gradient computed by the Hilbert transform, or selecting the frequency bin with the greatest energy in a real-time windowed FFT. In this implementation, a second order adaptive IIR lattice notch filter $H(z)$ is employed to track and remove the fundamental frequency with transfer function

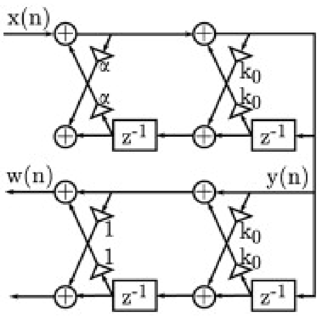
$$H(z) = \frac{W(z)}{X(z)} = \frac{1 + 2k_0[n]z^{-1} + z^{-2}}{1 + k_o[n](1+\alpha)z^{-1} + \alpha z^{-2}}, \quad W(z) = \sum_{k=-\infty}^{\infty} w[k]z^{-k} \tag{12}$$

where $W(z)$ is the Z-transform of the upper second section output w[n] shown in Figure 5(a), x[n] is a single channel of the input signal downsampled by factor M, and α controls the notch bandwidth. The lattice coefficient $k_0[n]$ is bounded by defining it as the sigmoid function

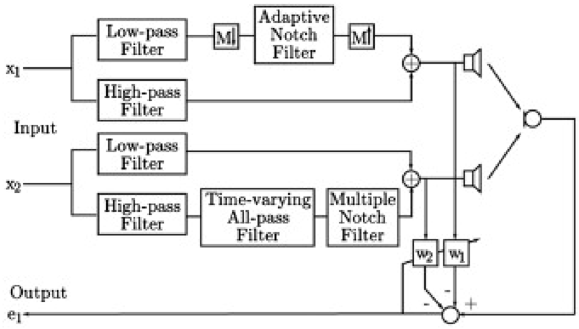
$$k_0[n] = \frac{2}{1 + e^{-g_0[n]}} \tag{13}$$

such that $|k_0[n]| < 1$ and updated using the minimization rule

$$\nabla_{g_0}\left(\sum_{k=0}^{n} \lambda^{n-k} w^2[k]\right) = 0, \quad 0 < \lambda < 1 \tag{14}$$



(a) IIR lattice notch filter diagram



(b) Decorrelation algorithm block diagram

where λ is a chosen forgetting factor. The estimated fundamental frequency $\hat{f}_0[n]$ can then be found by knowledge of $k_0[n]$ as

$$\hat{f}_0[n] = \frac{f_s}{2\pi M}\cos^{-1}(-k_0[n]) \tag{15}$$

where f_s is the sampling frequency. A time-varying second order all-pass filter $A(z)$ applied to an adjacent channel with transfer function

$$A(z) = \frac{k_0^2[n] - 2k_0[n]z^{-1} - z^{-2}}{1 - 2k_0[n]z^{-1} + k_0^2[n]z^{-2}} \tag{16}$$

introduces a group delay falling within the JND-ITD thresholds mentioned in Section 2.3 for each frequency to further improve decorrelation performance while maintaining the magnitude response due to its all-pass characteristic. In this case, the group delay across the frequency range falls under 40 μs, meaning the added phase distortion is imperceivable under regular listening conditions. The filter is also always stable and causal due to the boundedness of $k_0[n]$ as described earlier. Finally, a polynomial multiple notch filter is applied to the adjacent channel described by the transfer function

$$M(z) = \frac{\prod_{m=M_{min}}^{M_{max}}(1 - e^{j\omega_m(n)}z^{-1})}{\prod_{m=M_{min}}^{M_{max}}(1 - e^{j\omega_m(n)}\rho z^{-1})} \tag{17}$$

where $0 < \rho < 1$ is the pole radius and the notch frequencies

$$\omega_m(n) = 2\pi m\hat{f}_0[n], \quad 1 \leq M_{min} < m < M_{max} \leq \left\lfloor \frac{f_s}{\hat{f}_0[n]} \right\rfloor \tag{18}$$

are computed from $k_0[n]$ by the relation given in (15) between harmonics M_{min} and M_{max}. Choosing M_{min} to create the lowest harmonic frequency notch within the high frequency range of the signal is motivated by the perceptual insensitivity of phase and magnitude distortion at very high frequencies as seen in the JND-dB and JND-Hz curves in Figure 1.

Decorrelation performance between any two channels is measured by the magnitude square coherence (MSC) which is defined as

$$\mathrm{MSC}_{i,j}(f) = \frac{|S_{ij}(f)|^2}{S_{ii}(f)S_{jj}(f)}, \tag{19}$$

where $S_{ij}(f)$ is the cross-power spectrum between inputs channels x_i and x_j, and $S_{ii}(f)$ is the auto-power spectrum of a single channel x_i in the multichannel output. In this study, the case of $i, j = \{1, 2\}$, or stereo channel decorrelation, is assessed. Results for the stereo case of full-band input from [8] are presented in Figure 6.

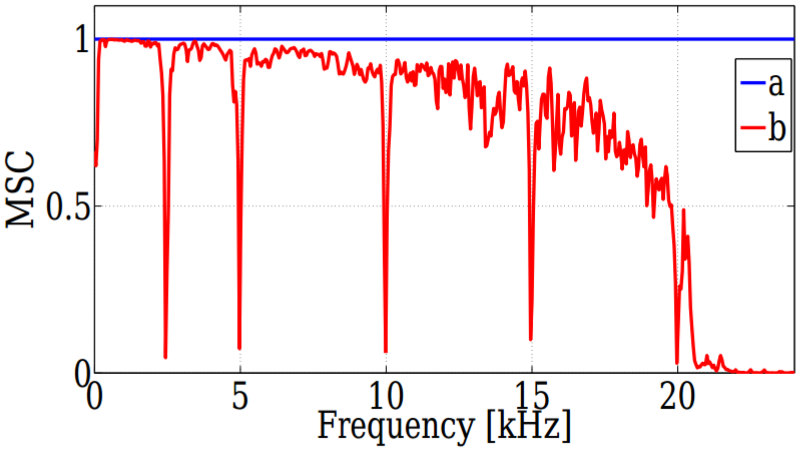


Figure 6: Magnitude square coherence before (a) and after (b) adaptive decorrelation method for $f_0 = 2.5$ kHz and harmonics 5 kHz, 10 kHz, 15 kHz, and 20 kHz

4 Conclusions

In this report, we have outlined the background, motivation, and formulation of practical implementations for two examples of personal sound zone generation and multichannel acoustic equalization using perceptually motivated adaptive filtering algorithms. By applying concepts from psychoacoustics, it is shown that the ability to improve on the subjective performance of existing filtering techniques which only consider optimization of physical parameters is possible. Likewise, the ability to efficiently improve performance through imperceptible filtering operations is also achievable. In the future, these implementations can be integrated within nonlinear approaches to filter generation involving deep neural networks (DNNs) and extended to scenarios involving head-tracking to further improve the capabilities of PSZ technology.

References

[1] Yue Qiao and Edgar Choueiri. Sann-psz: Spatially adaptive neural network for head-tracked personal sound zones, 2024.

[2] Malcolm Slaney. The 13 db miracle. `https://ccrma.stanford.edu/~malcolm/13dB_Miracle/`, 2024.

[3] Kyle Forinash and Wolfgang Christian. Just noticeable difference. `https://phys.libretexts.org/Bookshelves/Waves_and_Acoustics/Book%3A_Sound_-_An_Interactive_eBook_(Forinash_and_Christian)/07%3A_Pitch_Loudness_and_Timbre/7.01%3A_Pitch_Loudness_and_Timbre/7.1.04%3A_Just_Noticeable_Difference`, 2024.

[4] Weiping Tu, Ruimin Hu, Heng Wang, and Wenqin Chen. Measurement and analysis of just noticeable difference of interaural level difference cue. In *2010 International Conference on Multimedia Technology*, pages 1–3, 2010.

[5] Taewoong Lee, Jesper Kjær Nielsen, and Mads Græsbøll Christensen. Signal-adaptive and perceptually optimized sound zones with variable span trade-off filters. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2412–2426, 2020.

[6] Steven Par, Armin Kohlrausch, Heusdens Richard, Jesper Jensen, and Søren Jensen. A perceptual model for sinusoidal audio coding based on spectral integration. *EURASIP Journal on Advances in Signal Processing*, 2005, 06 2005.

[7] Stefania Cecchi, Laura Romoli, Paolo Peretti, and Francesco Piazza. Low-complexity implementation of a real-time decorrelation algorithm for stereophonic acoustic echo cancellation. *Signal Processing*, 92(11):2668–2675, 2012.

[8] Stefania Cecchi, Alberto Carini, Francesco Piazza, and laura romoli. A multichannel and multiple position adaptive room response equalizer in warped domain. 09 2013.