Source Separation with Perceptual Loss

Yifeng Yu, Alexander Lerch
Center for Music Technology, Georgia Institute of Technology, USA
{yyu479,alexander.lerch}@gatech.edu

***Abstract*—In the field of audio signal processing, source separation is a critical challenge with significant implications for applications such as music remixing, automatic transcription, and noise reduction. This challenge involves not only distinguishing different audio sources but also preserving their quality. Traditional approaches have primarily utilized Mean Squared Error (MSE) loss, focusing on mathematical precision. However, such methods do not necessarily align with human auditory perception. This paper addresses these limitations by proposing a novel approach that emphasizes perceptual quality over mathematical exactitude. By integrating psycho-acoustic principles, our proposed methods aim to produce separations that sound more natural and are less affected by artifacts, even if they allow for greater mathematical deviations compared to traditional MSE-based methods. Through various implementation strategies of perceptual loss, our research seeks to bridge the gap between quantitative metrics and qualitative listening experience, significantly enhancing the practical applications of audio source separation.**

I. Research Statement/Problem

SOURCE separation involves separating an audio signal into its distinct elements, such as individual instrument tracks and vocal lines. This task is particularly challenging in the realm of music production, automatic transcription, and noise reduction, where these elements often overlap both temporally and spectrally, creating complex patterns that are difficult to disentangle with precision. The purpose of this task is not only to enable and enhance activities such as music remixing and sampling but also to provide the foundation for improved music information retrieval, automatic transcription, and even noise reduction. However, the existing methods for source separation often fail to produce high-quality separations, especially in real-world audio scenarios where signals may be noisy and overlapping. Therefore, there is a need for an improved source separation method that can produce high-quality separations in such scenarios. Besides, a major drawback of current source separation methods is their reliance on metrics focused primarily on numerical accuracy, overlooking perceptual quality. This leads to systems that, while achieving low error rates as dictated by conventional loss functions and metrics, may fail to produce outputs that align with human auditory perception. Therefore, a refined approach prioritizing perceptual fidelity is essential for achieving truly high-quality separation in complex auditory scenarios.

II. Motivation

Source separation is fundamental to a variety of applications, ranging from speech enhancement to the isolation of individual musical tracks, all of which involve extracting different sound sources from a mixed audio signal. The quality of source separation is influenced by several key factors, including the choice of processing domain (time or frequency), the nature of the input data, and the architecture of the separation model itself. Among these, the selection of an appropriate loss function during training is crucial. The traditional mean squared error (MSE) loss in frequency or time domain, which penalizes the squared difference between the predicted and true sources, has been a popular choice due to its simplicity and empirical performance. However, those losses have a notable limitation: they focus on reducing errors in a strictly mathematical sense, which might not always translate to perceptually better results for human listeners.

The human auditory system is complex and our understanding of how we hear sounds is based on psycho-acoustic principles. In tasks such as audio source separation, numerical differences between generated and actual sounds can be substantial, yet these discrepancies may not always be perceptible to the human ear. On the other hand, small numerical differences can sometimes be very obvious and uncomfortable to listeners. This divergence between numerical precision and perceptual relevance highlights a fundamental limitation when using Mean Squared Error (MSE) as the primary loss function for tasks related to auditory processing. Such a function might inadequately represent how humans perceive sound, leading to outcomes that, while numerically accurate, do not align well with human auditory experience.

Therefore, inspired by the nuances of human auditory perception [1], we propose the development of a novel type of loss function called perceptual loss. By focusing on perceptual qualities instead of strict mathematical deviations, this loss function aims to produce results that are more aligned with how humans perceive sound. Intuitively, an algorithm optimized with perceptual loss might produce results that sound more natural and less artifact-ridden, even if it permits larger mathematical deviations compared to MSE.

Incorporating perceptual loss into source separation algorithms might bridge the gap between mathematical accuracy and human auditory perception. This approach may result in a trade-off where these algorithms potentially exhibit lower performance in traditional quantitative evaluations. However, the priority is to achieve a more authentic auditory experience, thereby enhancing qualitative feedback from listeners. The success of these algorithms is thus measured more by their perceptual fidelity to human listeners than by conventional numerical metrics. Given the myriad ways in which perceptual loss can be implemented, this research aims to delve into a variety of these strategies. By exploring these diverse approaches, we hope to identify methods that not only maintain mathematical precision but also closely mirror the intricacies

of the human auditory experience.

III. Related Work / Context

Source separation is a well-studied problem in the field of audio signal processing. Over the years, many different methods have been proposed to tackle this problem, ranging from traditional signal processing techniques to more recent deep learning approaches. One traditional method is non-negative matrix factorization (NMF), which has been used to separate audio signals by assuming that the sources are non-negative linear combinations of basis vectors [2]. Another method is independent component analysis (ICA), which aims to separate sources by finding a linear transformation that maximizes their statistical independence [3]. Recently, deep learning methods such as convolutional neural networks (CNNs) [4]–[6] and recurrent neural networks (RNNs) [7], [8] have been applied to source separation with great success. Another notable approach is the Wave-U-Net [9], [10], which operates directly on the waveform level, bypassing the need for spectrogram-based representations. This model has been shown to be particularly effective for music source separation. Recent developments have seen the integration of generative models, like Generative Adversarial Networks (GANs) [11], [12], into source separation tasks, pushing the boundaries of performance further. Additionally, the use of perceptual loss functions for neural modeling of audio systems, as explored in [13], delves into the implementation of perceptually relevant pre-emphasis filters in loss functions for audio processing neural networks. By incorporating lowpass filtering at high frequencies, it aims to enhance the perceptual similarity of model outputs to target devices, as confirmed by listening tests. This method underscores the growing emphasis on perceptual quality in audio signal processing, aligning model outputs more closely with human auditory perception.

IV. Method

A. Masking Thresholds Computation

Inspired by perceptual encoding [1], we propose a loss function based on the masking threshold of audios, reminiscent of mechanisms within MP3 encoders. These encoders, grounded in psycho-acoustic principles, prioritize human auditory perception by discerning which parts of an audio signal are most perceptible. Similarly, our method calculates the masking threshold across frequency bands, focusing on perceptually significant deviations between audio signals. By doing so, our approach ensures a more human-centric evaluation, capturing differences that align with genuine human auditory experiences. Here is the process of this algorithm:

1) **Fourier Transformation**: To begin with, we first transform our audio signal into its spectral domain using the Fast Fourier Transform (FFT). This step provides a frequency decomposition of the signals, which serves as a precursor to the subsequent processes. Let the time-domain signal be represented as $\boldsymbol{x}(t)$, then the Fourier Transform yields $\boldsymbol{X}(f)$, where f indicates the frequency components. For this transformation, we use an FFT size of 4096 and a hop size of 1024, with a sample rate of 44100 Hz.
2) **Bark Scale Mapping**: The human auditory system does not perceive all frequencies equally. To mimic this, we apply the Bark scale, a psycho-acoustic scaling of frequencies that mirrors the frequency resolution of the human ear. This transformation can be mathematically expressed as:

$$\text{Bark}(f) = 6 \cdot \arcsin\left(\frac{f}{600}\right) \tag{1}$$

where f is the frequency in Hz. By transforming our frequency representation $\boldsymbol{X}(f)$ into the Bark scale using this formula, we achieve a perception-aligned frequency decomposition, $\boldsymbol{X}_{\text{bark}}$. In our implementation, we define a set of 64 filter banks, designed to span the entire Bark scale up to the Nyquist frequency of the audio sampling rate. Each filter bank corresponds to a specific range on this scale. Frequencies are converted to their corresponding Bark values and then rounded to the nearest filter bank index. This transformation enables our model to better reflect the non-linear hearing sensitivity of human ears across different frequencies.
3) **Masking Threshold Computation**: The masking threshold represents the minimum audible threshold of sound in the presence of another, louder sound. By utilizing the masking threshold, we can focus our loss computation on differences that are perceptually significant to human listeners.
In our psycho-acoustic model, we implement a dynamic Bark-scale based masking function that models both upward and downward masking. The function is designed to simulate how certain frequencies can mask others, making them inaudible.
The spreading function in dB is calculated as follows:

$$\text{SF}_{\text{dB}}(x) = \begin{cases} -23.5 + -27 \cdot b_{\text{max}} + \frac{x \cdot (8 + 27 \cdot b_{\text{max}})}{N_{\text{bark}} - 1}, & x < N_{\text{bark}} \\ -23.5 + \frac{(x - N_{\text{bark}}) \cdot (-12 \cdot b_{\text{max}})}{N_{\text{bark}} - 1}, & x \geq N_{\text{bark}} \end{cases} \tag{2}$$

where b_{max} is the maximum Bark value derived from the Bark scale transformation formula, as shown in Equation 1, with f is set at 22050 Hz. The number N_{bark} refers to the number of Bark filter banks, set at 64.
This dB scale is then converted into a voltage scale by applying:

$$\text{SF}_{\text{Voltage}}(x) = 10^{\left(\frac{\text{SF}_{\text{dB}}(x)}{20} \cdot \alpha\right)} \tag{3}$$

where α is a scaling factor that can be adjusted based on the desired sensitivity of the masking effect, set at 0.8.
Once the spreading function $\text{SF}_{\text{Voltage}}(x)$ has been calculated, it is used in the convolution of $\text{SF}_{\text{Voltage}}(x)$ with $\boldsymbol{X}_{\text{bark}}$, which results in an array of masking thresholds $\boldsymbol{m}$ across the Bark scale. The convolution operation can be mathematically represented as follows:

$$\boldsymbol{m} = \text{SF}_{\text{Voltage}}(x) * \boldsymbol{X}_{\text{bark}} \tag{4}$$

4) **Adding Listening Threshold in Quiet (LTQ)**: Once the masking thresholds $\boldsymbol{m}$ are computed, an optional step involves applying the Listening Threshold in Quiet (LTQ) curve. The LTQ curve represents the minimal audible sound level for different frequencies under quiet conditions and is used to ensure that the masking thresholds do not fall below human auditory capabilities. The application of the LTQ curve is mathematically described as follows:
First, define the unbounded LTQ curve function as follows:

$$\begin{aligned}\text{LTQ}_{\text{raw}}(f) = 3.64 \cdot \left(\frac{f}{1000}\right)^{-0.8} \\ - 6.5 \cdot \exp\left(-0.6 \cdot \left(\frac{f}{1000} - 3.3\right)^2\right) \\ + 0.001 \cdot \left(\frac{f}{1000}\right)^4 \end{aligned} \tag{5}$$

This function quantifies the quiet threshold levels across frequency, but to ensure these levels remain within realistic and measurable limits, they are bounded using the following:

$$\text{LTQ}(f) = \max\left(-20, \min\left(120, \text{LTQ}_{\text{raw}}(f)\right)\right) \tag{6}$$

This ensures that the LTQ levels are constrained between -20 dB and 120 dB, reflecting the practical range of human auditory perception under quiet conditions.
The resulting LTQ values are used to update the masking thresholds, incorporating a reference decibel level, denoted as r, commonly set at 60 dB. The formula used is:

$$\boldsymbol{m}_{\text{LTQ}} = \max\left(\boldsymbol{m}, 10^{\frac{\text{LTQ}(f)-r}{20}}\right) \tag{7}$$

This converts the LTQ curve from a decibel scale to a magnitude scale, aligning it with the baseline auditory perception levels. By integrating these adjusted LTQ values with the original $\boldsymbol{m}$ values, the updated masking thresholds incorporate essential limits of human auditory sensitivity.

B. Perceptual Loss Design

1) **Original MSE loss**:
To establish a baseline for comparing the effectiveness of our perceptual loss function, we begin with the Mean Squared Error (MSE) loss. This traditional loss function is used in the frequency domain to quantify the difference between the ground truth audio spectrum $\boldsymbol{y}$ and the generated audio spectrum $\hat{\boldsymbol{y}}$, each consisting of N frequency bins. The MSE loss is defined as:

$$L_{\text{MSE}}(\boldsymbol{y}, \hat{\boldsymbol{y}}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - \hat{y}_i|^2 \tag{8}$$

Here, y_i and $\hat{y}_i$ represent the ground truth and generated audio spectra at the i-th frequency bin, respectively. N is the total number of frequency bins after FFT, which is set at 2049.

2) **LTQ-weighted Loss (LTQ-W)**:
In the previous work, the researchers employed three distinct types of filters applied to the time domain signal for loss calculations: a first-order highpass filter, a folded differentiator, and an A-weighting filter [13]. These filters were designed to modify the time-domain signals to reflect different aspects of auditory perception and processing before computing the loss.
In contrast, our method introduces an alternative way by employing the Listening Threshold in Quiet (LTQ) curve as a filter, but in the frequency domain. The LTQ curve, defined in Equation 6, adjusts the loss calculation based on the minimal audible thresholds across different frequencies, ensuring that our loss function is more aligned with human auditory sensitivity. The LTQ-weighted loss (LTQ-W) is formulated as follows:

$$L_{\text{LTQ-W}} = \frac{1}{N}\sum_{i=1}^{N}\left(|y_i - \hat{y}_i| \cdot 10^{\frac{-\text{LTQ}(f_i)}{20}}\right)^2 \tag{9}$$

Here, $\text{LTQ}(f_i)$ refers to the LTQ value for the frequency corresponding to the i-th bin, applied to adjust the significance of the error based on its audibility.

3) **Masking Thresholds Difference (MTD)**:
Building upon our understanding of psychoacoustic principles, we extend our approach to perceptual loss by incorporating the computed masking thresholds, which reflect the perceptual significance of various frequency components. The Masking Thresholds Difference (MTD) method evaluates the disparity in these thresholds between the ground truth and the generated audio signals, focusing specifically on how well the generated signals adhere to the psychoacoustic masking properties observed in the original signals. The Loss of MTD is defined as:

$$L_{\text{MTD}} = \frac{1}{N_{\text{bark}}}\sum_{i=1}^{N_{\text{bark}}}(m_i - \hat{m}_i)^2 \tag{10}$$

where m_i and $\hat{m}_i$ are the masking thresholds for the ground truth and generated signals, respectively, in the i-th Bark band, and N_{bark} is the total number of Bark bands considered, which is set at 64.

4) **Masking Thresholds Weighted Spectral Difference (MTWSD)**:
Expanding on this concept with an alternative approach that also leverages the insights gained from psychoacoustic principles, the Masking Thresholds Weighted Spectral Difference (MTWSD) method seeks to refine how perceptual accuracy is quantified. This method uses a loss function calculated from the MSE of the normalized differences between the predicted and ground true spectrograms, relative to the true masking threshold. The loss function is defined as:

$$L_{\text{MTWSD}} = \frac{1}{N}\sum_{i=1}^{N}\left(|y_i - \hat{y}_i| \cdot \min(2, \frac{1}{m_i + \epsilon})\right)^2 \tag{11}$$

The constant ϵ, set to $1e-8$, is included to ensure numerical stability and prevent division by zero. The use of the $\min(2, \frac{1}{m_i+\epsilon})$ function is particularly important as it prevents the loss from being excessively amplified when m_i is very small. By clamping the maximum weight value at 2, we ensure that the training process is more stable and converges more readily, avoiding erratic loss values that can impede the learning process.

To further explore performance in the decibel scale (dB), we have incorporated a logarithmic function into our loss calculation:

$$L_{\text{MTWSD-dB}} = \frac{1}{N}\sum_{i=1}^{N}\left(\log_{10}\left(\frac{|y_i-\hat{y}_i|}{m_i+\epsilon}+1\right)\right)^2 \quad (12)$$

This formulation leverages the logarithmic function to reflect the perceptual characteristics of human hearing. The addition of 1 inside the logarithm ensures that the function remains well-defined when the differences are close to zero, thereby avoiding the mathematical complications of taking the logarithm of zero.

5) **Scaled Masking Thresholds Weighted Spectral Difference (SMTWSD)**:
The perceptual loss function L_{SMTWSD} is defined as:

$$L_{\text{SMTWSD}} = \frac{1}{N}\sum_{i=1}^{N}\left(|y_i-\hat{y}_i|\cdot w_i\right)^2 \quad (13)$$

where w_i is the perceptual weighting defined as:

$$w_i = 1-\alpha+\alpha\cdot\max\left(\beta_{\text{min}}, \frac{\beta_{\text{max}}-m_i}{\beta_{\text{max}}}\right) \quad (14)$$

Here, β_{min} is a predefined minimum value set to 0.1, β_{max} is a predefined maximum value set to 7, α is a weighting parameter, set to 1 in default.

6) **Selective Audibility Loss (SA)**:
This loss function focuses on differentially weighting the errors between the predicted and ground truth audio spectra based on their audibility. The Selective Audibility Loss (SA) is defined as:

$$L_{\text{SA}} = \frac{1}{N}\sum_{i=1}^{N}\left(f(y_i,\hat{y}_i,m_i)\right)^2 \quad (15)$$

where the function $f(y_i,\hat{y}_i,m_i)$ is determined by:

$$f(y_i,\hat{y}_i,m_i) = \begin{cases} |\hat{y}_i-y_i|, & \text{if } y_i > m_i, \\ \max(0,\hat{y}_i-m_i), & \text{otherwise.} \end{cases} \quad (16)$$

The design of L_{SA} is based on the concept of masking thresholds that sound below these thresholds are likely to be masked and thus inaudible. Based on this principle, and referring to the scenarios depicted in Figure 1, there are four distinct cases to consider in our evaluation:

- When the true amplitude y_i exceeds the masking threshold m_i (Case 1 and Case 3), the loss is calculated as the absolute difference, emphasizing the errors that significantly impact the perceived audio quality.

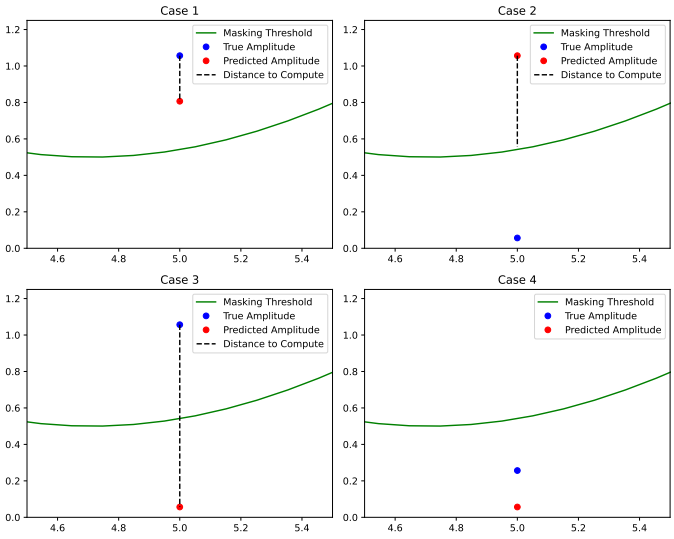

Fig. 1. Selective Audibility Loss Illustration, Different Positions of Predicted Values and Ground Truth Values in 4 Cases.

- If y_i is below m_i (Case 2 and Case 4), then the discrepancy is presumed inaudible. Thus, we only need to calculate the difference between $\hat{y}_i$ and m_i when $\hat{y}_i$ is above m_i; and the loss is 0 if both y_i and $\hat{y}_i$ are below m_i.

This approach ensures that the model focuses on errors that have a perceptual impact, discounting numerically minor or inaudible discrepancies, thereby aligning the model's outputs more closely with human auditory perception. It differentiates between perceptually significant and insignificant errors, optimizing computational efforts and enhancing the perceptual quality of the audio processing task.

7) **Softplus Selective Audibility Loss (SSA)**:
To introduce a smoother transition in error penalization at the audibility threshold, the Softplus Selective Audibility Loss (SSA) modifies the SA approach by employing a softplus function. This loss is formulated as:

$$L_{\text{SSA}} = \frac{1}{N}\sum_{i=1}^{N}\left(g(y_i,\hat{y}_i,m_i)\right)^2 \quad (17)$$

where the function $g(y_i,\hat{y}_i,m_i)$ is defined as:

$$g(y_i,\hat{y}_i,m_i) = \begin{cases} |\hat{y}_i-y_i|, & \text{if } y_i > m_i, \\ \text{softplus}(\hat{y}_i-m_i), & \text{otherwise.} \end{cases} \quad (18)$$

The softplus function is defined as:

$$\text{softplus}(x) = \log(1+e^x) \quad (19)$$

The softplus function smooths the transition when $\hat{y}_i$ is below m_i and ensures a more continuous sensitivity to errors around m_i.

8) **Selective Audibility Loss in Decibel Scale (SA-dB)**:
We can also convert the magnitude scale values into the decibel scale. This method differs from Equation 12,

where we convert each magnitude value into decibels. The vector $\boldsymbol{z}$ represents these magnitudes, which could be from either the predicted or ground truth spectrograms, or the masking thresholds. The transformation to the decibel scale is performed as follows:

$$\mathrm{dB}(\boldsymbol{z}) = 20 \cdot \log_{10}(\boldsymbol{z} + 1) \tag{20}$$

Applying this transformation, we compute $z_{i,\mathrm{dB}}$ for each individual component, whether it is y_i (the ground truth magnitude), $\hat{y}_i$ (the predicted magnitude), or m_i (the masking threshold) to obtain their corresponding decibel values as $y_{i,\mathrm{dB}}$, $\hat{y}_{i,\mathrm{dB}}$, and $m_{i,\mathrm{dB}}$ respectively.
The Selective Audibility Loss in Decibel Scale (SA-dB) is then defined as:

$$L_{\text{SA-dB}} = \frac{1}{N} \sum_{i=1}^{N} \left(f(y_{i,\mathrm{dB}}, \hat{y}_{i,\mathrm{dB}}, m_{i,\mathrm{dB}}) \right)^2 \tag{21}$$

This approach ensures the loss function remains sensitive to perceptually significant differences in quieter audio segments, leveraging the logarithmic nature of human auditory perception.

9) **Softplus Selective Audibility Loss in Decibel Scale (SSA-dB)**:
Similarly, for the Softplus Selective Audibility Loss in Decibel Scale (SSA-dB), we use:

$$L_{\text{SSA-dB}} = \frac{1}{N} \sum_{i=1}^{N} \left(g(y_{i,\mathrm{dB}}, \hat{y}_{i,\mathrm{dB}}, m_{i,\mathrm{dB}}) \right)^2 \tag{22}$$

10) **SMR-Weighted Loss (SMR-W)**:
The SMR-Weighted Loss introduces a scaling factor that depends on the signal-to-mask ratio (SMR) of each frequency bin, emphasizing errors in parts of the spectrum that are more audible and critical to perceptual quality. It is computed as:

$$L_{\text{SMR-W}} = \frac{1}{N} \sum_{i=1}^{N} \left(|y_i - \hat{y}_i| \cdot \min(2, \frac{y_i}{m_i + \epsilon}) \right)^2 \tag{23}$$

Here, $\frac{y_i}{m_i+\epsilon}$ calculates the SMR, taking into account not just the masking thresholds but also the magnitude of the actual signal, providing a comprehensive measure of audibility. This differs from Equation 11, which primarily focused on the masking thresholds without considering the signal's intensity.

V. Experiments

Our experiments were designed to evaluate the effectiveness of the Open-Unmix model in the task of audio source separation, specifically for separating vocals and accompaniment tracks. The following subsections detail the dataset, model configuration, and training parameters used in our experiments.

A. Dataset: MUSDB18

MUSDB18 [14] is a benchmark dataset that has been widely adopted in the music source separation community. It comprises a diverse range of tracks, ensuring a broad representation of musical genres, instrumentation, and recording conditions. This dataset provides a mixture as well as individual stems for vocals, bass, drums, and other accompaniment. Notably, MUSDB18 is structured with a pre-split configuration, consisting of 100 songs for training and 50 songs for testing, which makes it particularly well-suited for evaluating the effectiveness of our source separation approaches.

B. Model Configuration

The Open-Unmix [8] model, a bi-directional LSTM architecture in the field of music source separation, was employed due to its simplicity and efficacy in the source separation task. We configured the model to separate two primary tracks: vocals and accompaniment.

C. Training Parameters

Training was conducted over a maximum of 1000 epochs with an early stopping mechanism to prevent overfitting and to optimize computational resources. The early stopping criterion was set with a patience of 140 epochs, meaning that the training would cease if no improvement in validation loss was observed for 140 consecutive epochs.

The model was trained using a batch size of 16. The initial learning rate was set to 0.001, with the Adam optimizer. The learning rate decay was governed by a patience parameter of 80 epochs and a decay factor (gamma) of 0.3, allowing the model to make finer adjustments in the later stages of training.

Furthermore, a weight decay regularization was implemented with a coefficient of 0.00001 to encourage the model to maintain smaller weights, thereby helping to prevent overfitting and improve generalization on unseen data.

We use $\max$ and $+1$ in the logarithmic term to ensure that the ranges of those losses are similar to that of the original MSE loss, thus preventing the need to scale or tune the learning rate.

VI. Evaluation

This section presents the evaluation of perceptual loss applied in audio source separation, emphasizing the effectiveness of the approach through comprehensive metrics such as the Signal-to-Distortion Ratio (SDR), Signal-to-Interference Ratio (SIR), Signal-to-Artifact Ratio (SAR), and Image-to-Spatial Ratio (ISR). These metrics collectively provide a robust measure of quality for separated audio components, specifically vocals and accompaniment.

To ensure a rigorous and statistically meaningful evaluation, each track within the test set underwent a detailed analysis where the median SDR, SIR, SAR, and ISR values for each frame were calculated first. Subsequently, for a comprehensive representation of the dataset's performance, the median of these median frame values was computed across all 50 tracks—effectively creating a median of medians.

TABLE I
EVALUATION OF DIFFERENT LOSS FUNCTIONS OF OPEN-UNMIX

Loss Type	Use $\boldsymbol{m}_{\text{LTQ}}$	ISR (dB) (↑)		SAR (dB) (↑)		SDR (dB) (↑)		SIR (dB) (↑)	
		Vocals	Accompaniment	Vocals	Accompaniment	Vocals	Accompaniment	Vocals	Accompaniment
Baseline MSE	-	11.99	19.75	5.88	12.90	6.04	**12.49**	13.79	18.69
LTQ-W	-	10.79	20.14	5.43	12.81	5.66	11.81	13.57	16.77
MTD	-	6.70	20.68	4.06	11.50	4.98	10.77	13.58	12.52
MTWSD	✗	10.51	20.49	5.41	12.77	5.72	12.13	13.66	16.06
MTWSD	✓	10.51	21.64	5.49	13.05	5.82	11.94	15.11	16.13
MTWSD-dB	✓	2.17	**27.08**	0.25	**13.64**	0.98	7.46	14.97	8.23
SMTWSD	✗	12.12	20.98	5.77	13.30	6.16	12.40	14.51	18.20
SMTWSD	✓	11.74	20.36	5.37	12.84	5.98	12.25	13.84	17.11
SA	✗	12.35	19.78	6.00	12.78	5.68	12.07	13.33	18.68
SA	✓	12.57	19.07	5.80	12.64	5.77	12.02	12.40	17.97
SA-dB	✗	10.66	21.93	5.27	13.05	5.66	12.21	14.75	15.71
SA-dB	✓	11.00	21.85	5.37	13.07	5.54	12.23	14.34	15.74
SSA	✗	12.45	20.45	6.03	13.18	**6.20**	12.25	14.08	18.51
SSA	✓	11.78	20.64	5.85	13.07	6.03	12.16	13.75	17.93
SSA-dB	✗	10.44	21.87	5.49	13.42	5.56	12.31	**15.40**	16.78
SSA-dB	✓	10.19	21.87	4.94	13.06	5.35	11.84	14.25	15.21
SMR-W	✗	**18.44**	9.41	**6.16**	9.22	2.45	8.20	3.51	**21.65**
SMR-W	✓	13.21	16.4	5.78	11.83	4.85	10.8	9.67	19.19

A. Results and Discussion

The evaluation of various loss functions for the Open-Unmix model, as summarized in Table I, demonstrates significant variability in performance across different metrics and configurations. Notably, the inclusion of $\boldsymbol{m}_{\text{LTQ}}$, as an optional choice detailed by Equation 7, offers an innovative alternative to traditional masking thresholds by incorporating the LTQ. This approach, however, revealed mixed results in terms of overall audio quality enhancement.

Specifically, the LTQ-W, MTD, MTWSD-dB, and SMR-W loss functions did not improve audio quality across the four evaluation metrics. These loss functions exhibited issues with convergence; some did not converge effectively, while others converged too rapidly. For instance, LTQ-W tended to overlook details in high frequencies, resulting in audio that appeared more blurred compared to the baseline. Similarly, MTD, which only uses 64 values for comparison, may not effectively capture the full spectral detail critical for high audio quality. Additionally, MTWSD-dB highlighted the challenges associated with converting the magnitude of the signal into decibels, as described in Equation 12, proving this approach ineffective. SMR-W did not perform as well as MTWSD when an additional y_i was added to the weighting factor.

In contrast, SA, SSA, and SMTWSD presented competitive results in SDR metrics for vocals. Notably, SSA without LTQ achieved the highest scores among all methods tested. Generally, the performance improved when LTQ was not applied to the masking thresholds. Moreover, calculating differences on the magnitude scale proved to be more effective than utilizing the dB scale, underscoring the importance of choosing the right scale for loss calculation.

Overall, these findings highlight the complexity of audio source separation and the design of loss functions. The observed variability across different configurations and metrics indicates that no single approach consistently outperforms others in every aspect, emphasizing the necessity for a nuanced understanding of each method's strengths and limitations. Future research should continue to explore these dynamics to refine loss function designs further and achieve more reliable and perceptually aligned audio separation outcomes.

VII. NOVELTY OF PROPOSED WORK

The challenge of aligning objective metrics with human perceptual quality continues to be a significant hurdle in the field of source separation. Traditional loss functions prioritize mathematical precision but often fail to capture the subtleties of human auditory perception. Our work introduces novel methodologies designed to bridge this gap. Unlike conventional approaches that use deterministic loss functions, our approach incorporates perceptual encoding strategies akin to those used in advanced audio codecs like MP3. We integrate psycho-acoustic principles, including the Bark scale, LTQ, and masking thresholds, to develop loss functions that align more closely with how humans perceive sound.

VIII. DELIVERABLES

The code for the implementation of perceptual loss functions is made publicly available on GitHub[1].

IX. CONCLUSION

This paper has explored the design of various perceptual loss functions, incorporating psycho-acoustic principles, such as LTQ, masking thresholds, and dB scales that are crucial to aligning with human auditory perception. While these advanced loss functions showed potential in certain metrics, a general improvement across all metrics was not achieved. The absence of significant perceptual differences observed and time constraints prevented the implementation of listening tests and objective perceptual quality assessments such as PEAQ, which might have provided further insights into the subtle perceptual impacts of each loss function.

[1]See the implementation of perceptual loss functions at https://github.com/jerryuhoo/Perceptual-Loss

REFERENCES

[1] T. Painter and A. Spanias, "Perceptual coding of digital audio," *Proceedings of the IEEE*, vol. 88, no. 4, pp. 451–515, 2000.

[2] D. Lee and H. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, pp. 788–91, 11 1999.

[3] S. Choi, A. Cichocki, H.-M. Park, and S.-Y. Lee, "Blind source separation and independent component analysis: A review," *Neural Inf. Process. Lett. Rev*, vol. 6, 11 2004.

[4] E. Vincent, "Musical source separation using time-frequency source priors," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 14, no. 1, pp. 91–98, 2005.

[5] N. Takahashi and Y. Mitsufuji, "Multi-scale multi-band densenets for audio source separation," in *2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*. IEEE, 2017, pp. 21–25.

[6] A. Jansson, E. J. Humphrey, N. Montecchio, R. M. Bittner, A. Kumar, and T. Weyde, "Singing voice separation with deep u-net convolutional networks." in *ISMIR*, S. J. Cunningham, Z. Duan, X. Hu, and D. Turnbull, Eds., 2017, pp. 745–751. [Online]. Available: http://dblp.uni-trier.de/db/conf/ismir/ismir2017.html#JanssonHMBKW17

[7] S. Uhlich, M. Porcu, F. Giron, M. Enenkl, T. Kemp, N. Takahashi, and Y. Mitsufuji, "Improving music source separation based on deep neural networks through data augmentation and network blending," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2017, pp. 261–265.

[8] F.-R. Stöter, S. Uhlich, A. Liutkus, and Y. Mitsufuji, "Open-unmix - a reference implementation for music source separation," *Journal of Open Source Software*, vol. 4, no. 41, p. 1667, 2019. [Online]. Available: https://doi.org/10.21105/joss.01667

[9] D. Stoller, S. Ewert, and S. Dixon, "Wave-U-Net: A Multi-Scale Neural Network for End- to-End Audio Source Separation," in *Proceedings of the 19th International Society for Music Information Retrieval Conference*. ISMIR, Nov. 2018, pp. 334–340. [Online]. Available: https://doi.org/10.5281/zenodo.1492417

[10] T. Nakamura and H. Saruwatari, "Time-domain audio source separation based on wave-u-net combined with discrete wavelet transform," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 386–390.

[11] V. Narayanaswamy, J. Thiagarajan, R. Anirudh, and A. Spanias, "Unsupervised audio source separation using generative priors," *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*, vol. 2020-October, pp. 2657–2661, 2020.

[12] S. Joseph and R. Rajan, "Cycle gan-based audio source separation using time–frequency masking," *Circuits, Systems, and Signal Processing*, vol. 42, no. 2, pp. 1163–1180, 2023.

[13] A. Wright and V. Välimäki, "Perceptual loss function for neural modeling of audio systems," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 251–255.

[14] Z. Rafii, A. Liutkus, F.-R. Stöter, S. I. Mimilakis, and R. Bittner, "The MUSDB18 corpus for music separation," Dec. 2017. [Online]. Available: https://doi.org/10.5281/zenodo.1117372