Audio Engineering Society
Convention Express Paper 223

Presented at the 156th Convention
2024 June 15-17, Madrid, Spain

User preference evaluation of the masking ratio in multiple speaker scenarios

Xiaojing Liu[1] and Joshua Reiss[1]

[1]*Centre For Digital Music, Queen Mary University of London.*

Correspondence should be addressed to Xiaojing Liu (xiaojing.liu@qmul.ac.uk)

ABSTRACT

The masking effect among multiple audio tracks can lead to decreasing of audio clarity. Adjusting the masking ratio is one method to minimize the masking effect between multiple tracks. This paper presents a subjective experiment aimed at exploring user preferences for the Masker-to-Signal Ratio (MSR) in multi-speaker content. Through the subjective listening test, this study addresses three research questions regarding user preference and transfers the results of the experiment into corresponding LUFS values. Based on the experiment, our findings suggest that maintaining a MSR value less than -10 and a minimum loudness difference of approximately 14 LUFS, between the target track and other tracks is necessary to preserve the prominence of the target track. Additionally, a MSR close to 0, coupled with a loudness difference of around 10 LUFS, will improve clarity between the target track and multiple tracks. In the investigation, nearly half of the participants maintained a positive attitude towards multi-track audio.

1 Introduction

In multiple-speaker scenarios, a common challenge is the difficulty in understanding the content of multiple speakers due to competing voices and background noise. For instance, in virtual conferences or online group discussions, other speakers and ambient sounds can drown out the current speaker's voice, causing comprehension issues. To address this, one method is to minimize the auditory masking of the non-prominent signals [1]. This ensures that each speaker's voice stands out clearly, improving overall communication effectiveness.

Auditory masking is a common acoustic phenomenon where the perception of one sound is weakened in the presence of another sound. In a multi-channel environment, factors such as frequency overlap [2], phase relationships [3], and dynamic variations [4] make the masking effect more complex. In multiple track scenarios, numerous works and efforts have been made to enhance the user experience in audio quality. During audio mixing, audio engineers employ various techniques and tools to adjust the volume, timbre, and quality of these tracks to ensure that audiences can clearly perceive these key elements throughout the mixing process and enjoy them without being overshadowed by other sounds.

Masking ratio is a critical factor that determines the extent to which one audio track is masked by others in a multi-track environment. Adjusting the masking ratio aims to alleviate or resolve the masking effect between audio signals, thereby enhancing the clarity and intelligibility of the audio [5]. Masking ratio has been widely used in the field of audio coding and audio quality assessment. In Moving Picture Experts Group (MPEG) layer II [6], the masking ratio refers to the signal-to-masker ratio (SMR). The SMR measurement compares the energy of an input signal to a masking threshold. A positive SMR indicates the signal is audible, while a negative SMR means the signal is masked by other components and thus inaudible. In the work of Brandenburg [7], the first step is to calculate the error between a distorted signal and a reference signal. Then, it estimates the masking threshold from the reference signal and computes the

noise-to-mask ratio in each time frame using the Bark scale. The masking ratio metric has been incorporated into recommendation BS.1387 for the perceptual evaluation of audio quality (PEAQ) [8]. In the work of [9], they suggest a measurement method for cross-adaptive signal-to-masker ratio. The measurement methods are related to the ERB filter to calculate the critical bands based on the distribution of activity evoked by a sound along the basilar membrane, which is called the excitation pattern [10].

While previous research has focused on measuring the masking ratio in audio quality or using the masking ratio to improve the effectiveness of systems [5][11][12], there remains a conspicuous gap in understanding how these masking ratios influence user preferences and perceptual experiences in complex, multi-speaker scenes. In the model of P. Aichinger et al [13], the masking ratio relates to the overall loudness without a masking signal and the overall loudness with a masking signal present. They recommend a Masked-to-Unmasked-Ratio of at least 10% between the target instrument and overmix, ensuring the target signal has adequate identification. However, they mention that different scenarios and research objectives may require different values for the Masked-to-Unmasked-Ratio.

Understanding user preferences concerning masking ratios in multiple speaker scenarios can ensure an immersive and satisfying listening experience across diverse contexts. By elucidating these preferences, our goal is to contribute to the advancement of audio processing techniques and technologies aimed at enhancing clarity and intelligibility in multi-speaker audio environments. In this context, we pose three research questions:

Question 1*: How does the preferred masking ratio for the prominent signal vary across different multitrack scenarios?*

In this question, we aim to measure under what conditions the prominent signal can be easily heard. This approach helps us determine the circumstances where listeners can most easily discern and comprehend the prominent signal.

Question 2*: What is the preferred masking ratio to ensure the prominent signal is audible while still allowing other tracks to be heard clearly?*

Within the context of multi-person free discussion or debates [14], this question serves to mitigate the risk of information loss by preserving the clarity and intelligibility of essential signals amidst the complexity of discourse.

Question 3*: How do people feel when they hear multiple tracks?*

When individuals encounter multiple tracks, their perceptions are influenced by different factors such as individual hearing characteristics, perception habits, and listening preferences. This question aims at gauging subjective attitudes towards multi-track audio experiences.

This paper presents an experiment that explores user preferences regarding the masking ratio of multiple speaker contents and different numbers of tracks. Participants are asked to determine their preferred value of the masking ratio for various scenarios. Following the experiment, each participant is asked to identify the feelings influenced by the multi-track content. Through a combination of subjective user evaluations, this research endeavors to address three research questions concerning the masking ratios and relative levels for different scenarios. In this manner, it purposes to contribute to the progress of audio engineering practices and the optimization of multi-speaker audio systems, thus facilitating future research on speech enhancement systems.

2 Methods

Masking in a multichannel environment becomes more complicated, as audio clarity needs to consider not only loudness differences (LD) but also factors such as phase, frequency, temporal characteristics, and frequency resolution. An existing auto-mix system [15] utilizes cross-adaptive masking metrics and audio effects to minimize masking and increase perceived clarity for overall mixing. Through different audio effects, it adjusts the Masking-to-Signal Ratio (MSR) in terms of phase, frequency, and loudness. To adjust the MSR in different scenarios and achieve our research objectives, we propose a new prominent track enhancement system. This system is based on the existing minimization of the masking system [15]. This system enables the reduction of masking while allowing for quantification and adjustment of the masking ratio. To achieve this goal, we changed the objective function used in [15].

Objective Function

Following [15]'s work we get the masking value of each track. In equation (1), $M_{DO}(\mathbf{x}_C)$ is the sum of the masking value difference value of prominent track and other tracks. $\mathbf{x}_C$ is the vectorized representation of the system parameter control explained in [15]. A is the total number of tracks. $M_B(\mathbf{x}_C)$ is that the masking value of the prominent track (target signal). $M_i(\mathbf{x}_C)$ represents the masking value for the i track.

$$M_{DO}(\mathbf{x}_C) = \sum_{\mathrm{i}\neq\mathrm{B}}^{A}\left(M_B(\mathbf{x_C}) - M_i(\mathbf{x}_C)\right) \quad (1)$$

$M_d(\mathbf{x}_C)$ in the equation (2) is to minimize the difference in masking levels among all tracks except the prominent track.

$$M_d(\mathbf{x}_C) = \max\left(\|M_i(\mathbf{x_C}) - M_j(\mathbf{x_C})\|\right)$$

$$For\ i = 1, \dots, n; j = 1, \dots, n; i \neq j, i \neq B, j \neq B \quad (2)$$

In the end, we add total masking value into the system. $M_T(\boldsymbol{x}_C)$ is the sum of all tracks in [15]. w is a weighting value to adjust MSR in the subjective listening test.

$$\boldsymbol{x}_C = min_{x_C}(M_T(\boldsymbol{x}_C) + (M_{Do}(\boldsymbol{x}_C) * \boldsymbol{w}) + M_d(\boldsymbol{x}_C) \quad (3)$$

3 Experiment

Stimuli

The stimuli for the listening test were sourced from the AIM dataset [16], LibriSpeech [17], and multiple-speaker videos on YouTube. For the subjective listening test, we prepared six groups of stimuli, each with a different number of tracks, all of which are described in Table 1. Table 2 provides comprehensive data, including MSR values, weighting values (explained in equation 3), loudness of each track, and maximum LD between prominent tracks. The MSR value represents the ratio between the prominent track and rest tracks. All MSR values are calculated from [15]'s masking metrics. Some values in the table are highlighted in red to indicate instances where the loudness of the prominent track is smaller than other tracks. Loudness units are expressed in Loudness Units Full Scale (LUFS).

Scenario	Track	Description
1	3	One male and two females were selected from the AIM dataset [16] EN2009b Headset
2	3	One male and two females, all sourced from the AIM dataset EN2006b Headset
3	4	Two female and two male sounds, all sourced from the AIM dataset EN2002a Headset
4	4	All male sounds were cropped and redesigned from YouTube and the LibriSpeech dataset [17].
5	5	Three female and two male sounds were cropped and redesigned from YouTube
6	5	Three female and two male sounds sourced from the AIM dataset EN2001e Headset.

Table1. Stimuli description.

Scenario	MSR	Weighting Value	Max LD	Prominent Track	Track 1	Track 2	Track 3	Track 4
	10.08	0.005	4.48	-20.81	-20.29	-25.29		
	-0.89	0.15	7.95	-21.15	-26.71	-29.10		
Scenario 1	-0.45	0.3	8.92	-18.24	-26.16	-27.16		
	-1.46	0.6	10.27	-14.22	-22.62	-24.48		
	-5.09	1.2	11.53	-14.73	-22.94	-26.26		
	11.37	0.005	0.67	-24.05	-23.54	-23.38		
	0.66	0.15	4.14	-22.54	-26.68	-25.59		
Scenario 2	-3.87	0.3	8.10	-19.90	-28.00	-24.54		
	-0.42	0.6	8.93	-18.65	-27.59	-25.17		
	-1.64	1.2	10.46	-18.75	-24.31	-29.21		
	5.68	0.005	1.43	-24.57	-24.85	-23.14	-25.55	
	-3.98	0.15	6.18	-21.38	-25.57	-24.95	-27.57	
Scenario 3	-7.55	0.3	9.24	-17.12	-21.84	-24.13	-26.36	
	-11.05	0.6	12.76	-13.61	-26.37	-23.73	-26.22	
	-14.15	1.2	18.87	-13.35	-32.22	-24.77	-26.52	
	15.03	0.005	1.03	-24.82	-25.17	-25.62	-25.85	
	3.49	0.15	4.25	-23.22	-27.02	-27.48	-26.63	
Scenario 4	-12.88	0.3	9.62	-18.19	-26.56	-27.81	-26.30	
	-17.94	0.6	13.26	-15.64	-27.56	-26.84	-28.90	
	-21.64	1.2	15.74	-14.86	-24.09	-30.60	-25.71	
	29.03	0.005	3.23	-23.95	-25.51	-21.70	-27.19	-20.83
	28.92	0.15	7.32	-27.23	-27.34	-19.92	-26.57	-24.65
Scenario 5	26.60	0.3	14.21	-17.13	-25.98	-18.64	-29.30	-31.33
	26.05	0.6	11.36	-15.84	-23.57	-20.19	-27.20	-27.16
	25.98	1.2	14.71	-15.57	-30.29	-17.22	-26.18	-23.30
	0.13	0.005	7.86	-28.73	-20.87	-29.07	-28.00	-21.00
	1.32	0.15	6.13	-30.18	-24.05	-26.95	-29.33	-25.30
Scenario 6	2.26	0.3	2.43	-25.82	-24.41	-25.83	-28.25	-27.40
	-1.37	0.6	8.20	-19.08	-15.30	-27.28	-24.30	-22.29
	-11.71	1.2	13.50	-14.43	-22.86	-25.84	-27.33	-27.93

Table2. The MSR, Weighting value, LD, and Loudness (LUFS) of each track.

Subjective listening Test

During the listening test, we utilized the Go Listen platform [18] to conduct a blind comparison test. In the listening test, we explore three research questions. For questions 1 and 2, participants were instructed to listen to the reference soundtrack initially. The reference soundtrack is a sole prominent track without processed by the system. Subsequently, participants were asked to listen to five complete soundtracks (including multiple tracks with the prominent track) generated by the system with varying weighting values.

Question 1

In this question, we inquired participants whether they could easily identify the reference signal in multiple tracks. The research question aims to assess the impact of the MSR on participants' ability to recognize reference signal in multiple tracks.

Question 2

In this question, we asked participants whether they could easily identify the reference signal while simultaneously recognizing other audio tracks. The research question also aims at assessing the impact of MSR on participants' perception when identifying the reference signal and other audio tracks.

Question 3

After the listening test, participants were asked: "Do you want to hear multiple sounds?". This inquiry seeked to gauge their preference and attitude for listening to multiple audio stimuli simultaneously and to understand their willingness and feeling to engage in tasks involving multiple tracks.

Participants

A total of 11 participants joined the listening test. All the participants of the subjective listening test are experienced audio engineers or music producers. Participants are required to conduct the task in a quiet, noise-free environment, preferably in a soundproof room, and wear monitoring headphones.

4 Results

Question1

In the first experiment, participants were asked to rate the prominence (ease of hearing) of the reference signal. Six scenarios with different weighting values were selected for analysis. Figure 1 displays the mean scores with error bars representing the 80% confidence intervals using the T-distribution. Figure 1 shows that, in most scenarios, the ratings increased with higher weighting values, indicating that the weighting value affects the prominence of the reference signal. This trend of increasing ratings with higher weighting values is particularly evident in Scenario 3 and Scenario 4. However, Scenario 1 and Scenario 5 did not exhibit this trend. Overall, it can be observed that the highest scores were consistently associated with weighting values of 0.6 and 1.2.

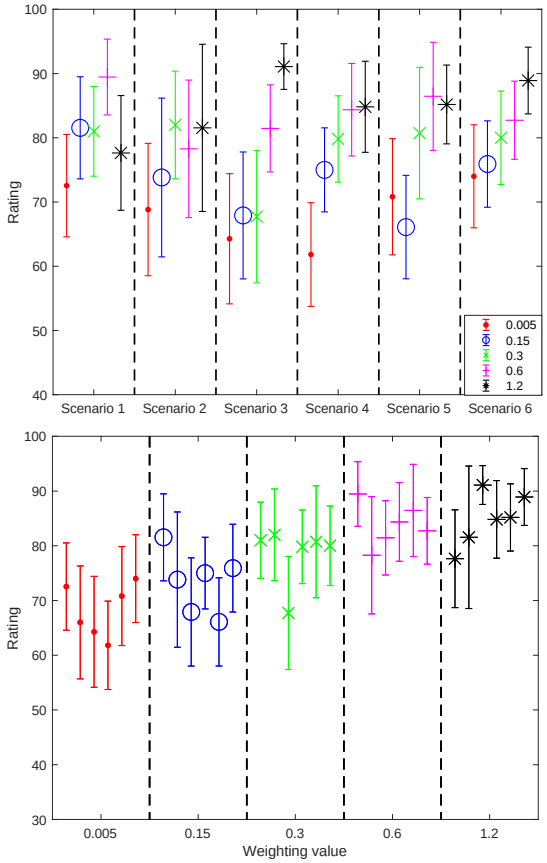


Figure1. Result of Question 1 *with 80% confidence intervals.*

Question2

In the second experiment, participants were directed to identify the reference signal while simultaneously discerning other audio tracks. Similar to the first experiment, analysis was conducted using six scenarios with varying weighting values. Figure 2 presents the mean scores, with error bars indicating the 80% confidence intervals using the T-distribution. Although the data points appear scattered, an overall trend emerges, showing that weighting values of 0.15 and 0.6 outperformed others in certain groups.

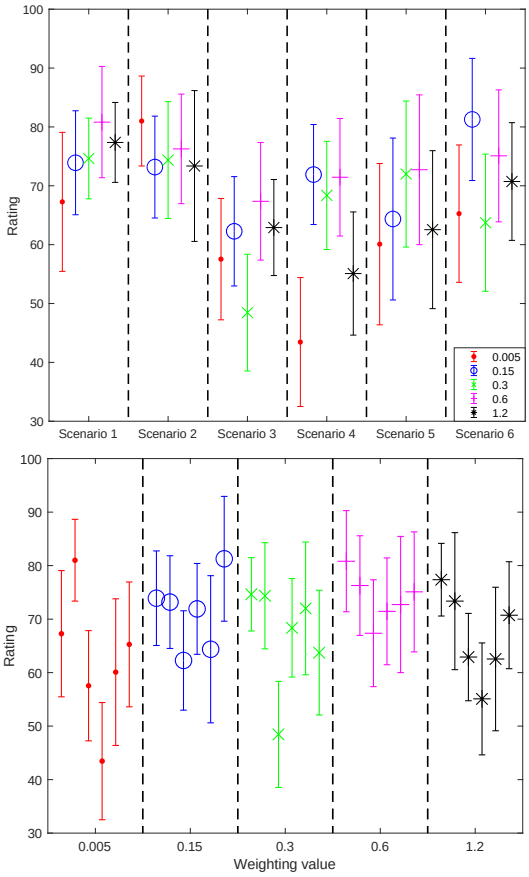


Figure2. Result of Question 2 with 80% confidence intervals.

Overall mean and median results for all stimuli are displayed in Figure 3 for systems varying weighting values. These give a clearer depiction of the overall scenarios of each weighting value, The 1.2 weighting value can be seen to perform best overall for Question 1, and the 0.6 weighting value for Question 2. Combining this fact with Table 2, the average LD of Question 1 is 14 LUFS, and the LD range is 14 ± 4 between the prominent track and other tracks for clarity to keep the target track prominent. In Question 2, the average LD is 10 LUFS, and the range of LD should keep around 10 ± 2 LUFS between the prominent track and other tracks for clarity. Figure 4 illustrates the relationship between average rating and MSR values. For Question 1, high rating values are observed when the value of MSR ranges between -10 and -20 (Combining with Table 2), with no occurrences of low scores. In Question 2, when the MSR value is close to 0, the rating value is high and there are almost no low scores.

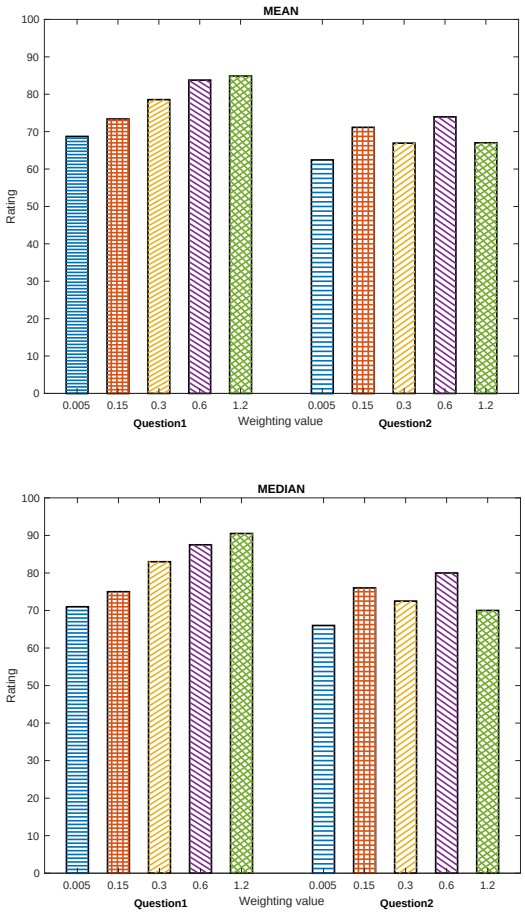
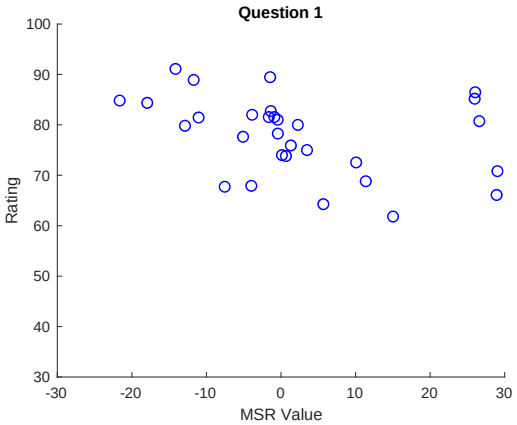


Figure3. Overall mean and median results for both experiments.



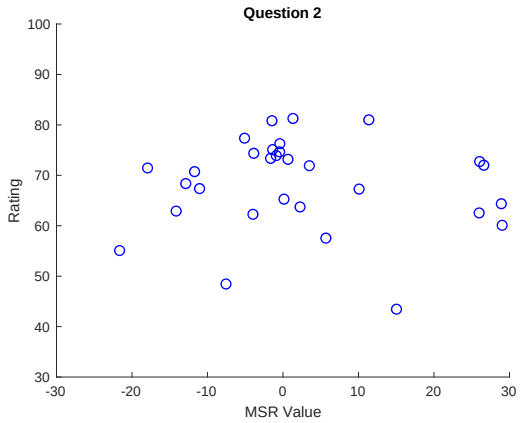


Figure4. Overall mean and MSR results for both experiments

Question3

In the 11 participants, 4 people pointed out that they did not want to hear multiple sounds. 2 people chose "other". 5 people thought the multiple sounds would not affect the perception.

5 Discussion

In this experiment, we formulated three research questions and conducted a web-based subjective listening test to investigate user preferences regarding the MSR in multi-speaker content. Our results reveal that for Research Question 1, maintaining MSR value less than -10, LD of around 14 LUFS between the prominent track and other tracks, is crucial to ensure the prominence of the target track. Furthermore, increasing the weighting value of the mixing system enhances the prominence of the target signal. For Research Question 2, we found that LD of around 10 LUFS between the prominent track and other tracks, along with the MSR value close to 0, allowing the prominent signal to remain audible while ensuring clear perception of other tracks.

Concerning Research Question 3, we discovered that nearly half of the participants held the belief that multiple sounds would not affect their perception. Nevertheless, in post-listening test interviews, certain participants articulated a reluctance towards multitrack audio, attributing it to perceived interference with auditory perception and noting a degree of discomfort. However, one participant provided a contrasting viewpoint, asserting that multitrack audio does not impede auditory perception. The participant highlighted the prevalent shift to remote work post-COVID-19, where online meetings often host more than 20 individuals in a single digital space. Consequently, he has adapted to listening to conversations or deciphering mumbles in multitrack environments. Moreover, several participants exhibited interests in the potential of multitrack communication systems. They noted the current communication devices rely on stereo and advised that with the increasing prevalence of surround sound technologies, multitrack communication could emerge as a significant trend in the future. Additionally, another participant proposed that integrating visual cues with scenarios involving multiple speakers could enhance the distinction between individual speakers.

Limitation

The current system is non-linear, non-convex, and utilizes integer optimization for parameters in harmony searching [15]. The entire system is influenced by input parameters or starting points. Consequently, we have limited flexibility in choosing the weighting value. Future work requires more refined adjustments to the MSR. The current paper only analyzes stimuli and results based on LD and masking ratio. Future work can further evaluate spatial quality, speech quality, and frequency variations.

6 Conclusion

Our study explores user preferences regarding masking ratios in multiple speaker scenarios. Through our experiments, we assess participants' preferences when exposed to various masking ratios for different research goals. Our findings indicate that the LD of approximately 14 LUFS between the prominent track and other tracks, along with the MSR below -10, is crucial to ensure the prominence of the target track. The MSR value close to 0, coupled with the LD of around 10 LUFS between the prominent track and other tracks, allows the prominent signal to remain audible while ensuring a clear perception of other tracks. Furthermore, around half of the participants expressed positive attitudes towards multiple speaker scenarios.

7 Acknowledgement

This research received support from the China Scholarship Council and Queen Mary University of London. I am grateful to Hongwei Ai and Dr. Shuren Tan for their insightful comments and suggestions, which greatly contributed to this study. Special thanks to all the participants who generously volunteered their time and provided valuable feedback for the listening tests. Their contributions were essential to the success of this research.

References

[1] A. Tsilfidis, C. Papadakos, and J. Mourjopoulos, "Hierarchical Perceptual Mixing," in Proc. 126th Audio Engineering Society Convention, Munich, Germany, May 7–10, 2009.

[2] R. P. Carlyon, "Encoding the fundamental frequency of a complex tone in the presence of a spectrally overlapping masker," in Journal of the Acoustical Society of America, vol. 99, no. 1, pp. 517–524, 1996.

[3] S. P. Bacon and D. W. Grantham, "Modulation masking: Effects of modulation frequency, depth, and phase," The Journal of the Acoustical Society of America, vol. 85, no. 6, pp. 2575–2580, Jun. 1989.

[4] B. Roberts, R. J. Summers, and P. J. Bailey, "Formant-frequency variation and informational masking of speech by extraneous formants: Evidence against dynamic and speech-specific acoustical constraints.," Journal of experimental psychology. Human perception and performance, vol. 40, no. 4, pp. 1507–1525, Aug. 2014.

[5] J. Nikunen and T. Virtanen, "Noise-to-mask ratio minimization by weighted non-negative matrix factorization," 2010 IEEE International Conference on Acoustics, Speech and Signal Processing, Dallas, TX, USA, pp, 25-28. 2010.

[6] K. Brandenburg and G. Stoll, "ISO/MPEG-1 audio: A generic standard for coding of high-quality digital audio," J. Audio Eng. Soc., vol. 42, no. 10, pp. 780–792, 1994

[7] K. Brandenburg and T. Sporer, "NMR and Masking Flag: Evaluation of Quality Using Perceptual Criteria", Proceedings of the AES 11th International Conference on Test and Measurement, pp. 169-179, May 1992.

[8] T. Thiede, W. C. Treurniet, R. Bitto, C. Schmidmer, T. Sporer, J. G. Beerends, et al., "PEAQ - The ITU Standard for Objective Measurement of Perceived Audio Quality", Journal of the Audio Engineering Society, vol. 48, pp. 3-29, 2000.

[9] S. Vega and J. Janer. Quantifying masking in multi-track recordings. In Proceedings of SMC Conference, 2010.

[10] B. R. Glasberg and B. C. Moore, "Derivation of Auditory Filter Shapes from Notched-Noise Data," Hearing Res., vol. 47, no. 1-2, pp. 103–138,1990.

[11] W. G. C. Bandara, N. Patel, A. Gholami, M. Nikkhah, M. Agrawal, and V. M. Patel, "Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders," in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14507-14517, 2023.

[12] C. Hummersone, T. Stokes, and T. Brookes, "On the ideal ratio mask as the goal of computational auditory scene analysis," in Blind source separation: advances in theory, algorithms and applications, Berlin, Heidelberg, Springer Berlin Heidelberg, 2014, pp.349-368.

[13] P. Aichinger, A. Sontacchi, and B. Schneider-Stickler, "Describing the transparency of mixdowns: The Masked-to-Unmasked-Ratio," 2011.

[14] H, Bo, et al. "Distributed System for Virtual Conference Audio Synthesis." IEEE Xplore, ieeexplore.ieee.org/abstract/document/540582,2009.

[15] X. Liu, A. Mourgela, H. Ai, and J. D. Reiss, "An automatic mixing speech enhancement system for multi-track audio," arXiv.org, Apr. 27, 2024. https://arxiv.org/abs/2404.17821

[16] J. Carletta et al., "The AMI Meeting Corpus: A Pre-announcement," in Machine Learning for Multimodal Interaction, 2006.

[17] H. Zen et al., "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," arXiv.org, arXiv link, 2019.

[18] D. Barry, et al. "Go Listen: An End-To-End Online Listening Test Platform." Journal of Open Research Software, vol. 9, 10.5334/jors.361. 2021.