X. Liu, H. Ai, and J. D. Reiss, "An Automatic Mixing Speech Enhancement System for Information Integrity," *J. Audio Eng. Soc.*, vol. 73, no. 10, pp. 648–659 (2025 Oct.). https://doi.org/10.17743/jaes.2022.0222.

An Automatic Mixing Speech Enhancement System for Information Integrity

XIAOJING LIU,[1,*] HONGWEI AI,[2] AND JOSHUA D. REISS,[1] *AES Fellow*
(xiaojing.liu@qmul.ac.uk) (hongwei.ai@gmail.com) (joshua.reiss@qmul.ac.uk)

[1] *Centre for Digital Music, Queen Mary University of London, London, UK*
[2] *Independent Researcher*

The simultaneous presence of multiple audio signals can lead to information loss due to auditory masking and interference, often resulting in diminished signal clarity. The authors propose a speech enhancement system designed to present multiple tracks of speech information with reduced auditory masking, thereby enabling more effective discernment of multiple simultaneous talkers. The system evaluates auditory masking using the ITU-R BS.1387 Perceptual Evaluation of Audio Quality model along with ideal mask ratio metrics. To achieve optimal results, a combined iterative Harmony Search algorithm and integer optimization are employed, applying audio effects such as level balancing, equalization, dynamic range compression, and spatialization, aimed at minimizing masking. Objective and subjective listening tests demonstrate that the proposed system performs competitively against mixes created by professional sound engineers and surpasses existing automixing systems. This system is applicable in various communication scenarios, including teleconferencing, in-game voice communication, and live streaming.

0 INTRODUCTION

In multiple audio processing systems, information loss is a critical issue, particularly when multiple sources transmit simultaneously, leading to difficulties in comprehension. In the early 1950s, air traffic controllers encountered significant challenges when managing communication [1]. Controllers had to listen to multiple pilots speaking over a single loudspeaker in the control tower, making it difficult to distinguish between the various voices and messages. This challenge is closely related to the phenomenon known as the "cocktail party effect."

Cherry reported the cocktail party effect in 1953 [2]. This phenomenon describes the human ability to focus on a specific sound or conversation while filtering out other sounds in a noisy environment, such as a restaurant or reception. Studies on auditory attention and selective hearing have deepened the understanding of the cocktail party effect and revealed how the brain distinguishes sound sources by analyzing the spatial localization and frequency characteristics of sounds [2]. Advances in neuroscience research have further allowed researchers to explore the brain's processing of multiple sound source environments. These studies have shown that the auditory cortex exhibits a high degree of dissociation and adaptability when processing information from multiple sound sources, helping to explain human auditory performance in complex acoustic environments like cocktail parties.

Researchers have explored the application of the cocktail party effect to address challenges such as source separation [3, 4], voice enhancement [5], and hearing loss [6, 7]. Some voice enhancement methods aim to extract a target speaker in a multispeaker environment[1] and reduce unwanted sources, environmental noise, or reverberation [8]. However, when the target speech signal involves multiple speakers, performance can be adversely affected by variations in speaker characteristics. Moreover, the majority of voice enhancement work focused on extracting target voice and ignoring other tracks that might include important information, thereby leading to information loss.

The goal of this paper is to develop a system capable of presenting multiple tracks of speech information, allowing users to concentrate on one particular track while seamlessly shifting their attention to another track as needed. This system leverages advanced signal processing algorithms to enhance the naturalness and intuitiveness of multisource speech interactions. By addressing the existing challenges related to information loss and attention man-

*To whom correspondence should be addressed, email: xiaojing.liu@qmul.ac.uk.

[1] In this paper, "speaker" refers to a human talker.

agement, this system aims to provide a more coherent and efficient auditory experience in complex acoustic environments.

This study proposes a lightweight system that leverages the Perceptual Evaluation of Audio Quality (PEAQ) model [9] to simulate human auditory perception and utilize adaptive ideal ratio masking metrics to assess auditory masking. The system effectively addresses both frequency and loudness masking, while also incorporating phase information. An iterative Harmony Search algorithm [10] is employed to optimize parameters related to audio effects, including equalization (EQ), dynamic range compression (DRC), spatialization (SPA), and level balancing. The range of these parameters are set according to established audio engineering practices. A web-based implementation of the system will be provided. Performance will be assessed through objective tests, and subjective evaluations will compare the proposed system against mixes created by professional sound engineers and existing automixing systems [11].

The rest of this research paper is organized as follows: SEC. 1 provides a brief overview of multiple speech enhancement techniques. SEC. 2 introduces the proposed method in detail, followed by SEC. 3, which discusses both the subjective and objective evaluations. Finally, SEC. 4 presents the conclusions of the paper.

1 BACKGROUND

1.1 Overview Of Speech Enhancement

Speech enhancement techniques can be mainly divided into single-channel (monophonic) and multichannel (stereophonic or multichannel) categories [12]. Single-channel speech enhancement techniques deal with audio signals captured by a single microphone. Multichannel speech enhancement involves deriving a clear speech estimate from multiple channels of mixed recordings [13]. Single-channel enhancement can be broadly categorized into three types: time-domain methods, frequency-domain methods, and time frequency-domain methods [14].

Multitrack enhancement involves a variety of audio processing techniques, which can mainly be divided into blind source separation (BSS) and beamforming approaches [15]. Beamforming is a technique that includes two main approaches: adaptive beamforming and directional beamforming [16]. Adaptive beamforming adjusts the weights of multiple microphone array channels to enhance signals coming from a specific direction while suppressing interference from other directions. In contrast, directional beamforming collects signals from a fixed direction, making it suitable for scenarios where the source direction is known [17].

Another multichannel speech enhancement technique is BSS. In BSS, Independent Component Analysis assumes that the sources are statistically independent and uses statistical methods to separate individual sources, while frequency-domain BSS processes signals in the frequency domain to improve computational efficiency and separation performance [18]. Furthermore, a successful method called neural beamforming combines supervised single-channel techniques with unsupervised beamforming for multiple speech enhancement [19, 20]. A neural network estimates the second-order statistics of speech and noise using time–frequency masks, and then a beamformer is used to linearly combine the multichannel mixture to produce clean speech. Besides beamforming and BSS, some researchers utilize spatial information of sound sources to improve the quality and clarity of speech signals.

1.2 Spatial Audio in Speech Enhancement

In the research of [21], researchers found that increasing the spatial separation between the signal and the masker enhanced the ability to reduce sound source or message uncertainty. Similarly, in multiple voice scenarios, Skowronek and Raake [22] investigated the impact of bandwidth, spatial audio reproduction, and communication complexity on user experience in multiparty conferencing. They conducted a subjective listening test using narrowband nonspatial (300–3,400 Hz), full-bandwidth nonspatial, and full-bandwidth spatial audio. The listening test results showed that the full-bandwidth spatial audio provided a statistically significant improvement over other conditions in terms of voice intelligibility, audio quality, attention, and user satisfaction. Notably, the spatial audio exhibited at least a 50% increase in these metrics compared to the narrowband nonspatial condition and a 25% increase compared to the full-band nonspatial condition. These findings highlight the benefits of spatialized audio in reducing cognitive load and enhancing perceived audio quality.

In multichannel surround sound systems, such as those used in television program mixing, dialogue is typically positioned at the front of the sound field to enhance prominence over other sound sources. Roginska and Geluso [23] explored the relationship between sound source positioning and audio clarity. Rothbucher et al. [24] further researched a teleconferencing system combining head-related transfer function with voice over Internet protocol, enabling online sound localization, separation, speaker detection, and channel allocation in conference call scenarios. The aforementioned works [21–24] demonstrated the value of spatial audio in enhancing user experience, particularly in teleconferencing applications. However, they did not address the issue of masking, such as how spatial audio could be used to reduce masking effects in complex audio scenes.

The work of [25] considered the microphone arrangement through spatial filtering. Their methods exploited time–frequency masking from multiple microphones in space and time to distinguish different sound sources, thereby improving the accuracy and effectiveness of speech enhancement. However, the research required extensive computing resources for training and inference. Furthermore, the speech signal was dynamic and spatial filtering would result in masking of other tracks. The system should consider changes in masking caused by changes in sound source location and how to deal with this change. The work of [11] suggested an automatic mixing system for audio clarity. The system uses a force-directed model to perform

SPA. In addition, loudness balancing and EQ are applied to ensure equal average perceptual loudness for each track across frequency bands. However, the equal average perceptual loudness in each frequency band might result in even more frequency masking.

While spatial audio techniques have shown promise in improving clarity and localization, they often fail to address masking effects caused by overlapping sound sources in multisource environments. To tackle this challenge, several researchers have employed masking techniques in speech enhancement.

1.3 Auditory Masking in Speech Enhancement

In the work of Heymann et al. [26], they considered a spectrum mask method based on the expectation maximization algorithm and the parameters of the Watson mixture model [27]. Jiang et al. [28] evaluated ideal binary masks in computational auditory scene analysis. They found that ideal binary masks had optimal performance in time–frequency units. Pfeifenberger et al. [29] suggested Eigennet architecture for estimating a gain mask metric. The system utilized spatial and amplitude information from power spectral density. Using a binary mask involves making a hard decision by either fully retaining or completely removing parts of the signal. However, this method can lead to the removal of background noise in time–frequency units where the speech is not prominent, which may ultimately reduce the overall hearing quality [30].

According to the aforementioned drawbacks, researchers have suggested replacing the ideal binary mask with the ideal ratio mask in [31], as it offers better noise suppression in challenging acoustic environments. In [32], it is posited that intertrack masking can significantly affect the overall clarity of the audio signal. As mentioned in [33], a cross-adaptive filter is able to handle input data from different sources or with varying characteristics, enhancing the model's robustness to various types of data.

One study [34] proposed an ideal ratio mask method, which utilized cross-adaptive masking metrics based on the Layer II metrics of the Moving Picture Experts Group (MPEG) [35] to reduce masking in multiple tracks. However, this work only considered frequency masking and ignored the masking of spatial position. Additionally, Hu et al. [36] implemented a basic masking threshold and compared the PEAQ [9] masking threshold curve with the MPEG Model II. They found that the PEAQ model more accurately characterized the auditory properties of the human ear while the MPEG Model II was more sensitive to the distribution of spectral energy.

In summary, spatial audio techniques have effectively improved clarity and localization, while masking-based methods like ideal binary masks excel in reducing frequency masking. However, few studies simultaneously consider the interactions between spatial and frequency masking, and evaluate in multispeaker environments.

To bridge this gap, the authors propose a novel system that systematically optimizes audio effects parameters to minimize unwanted masking and enhance overall clarity.

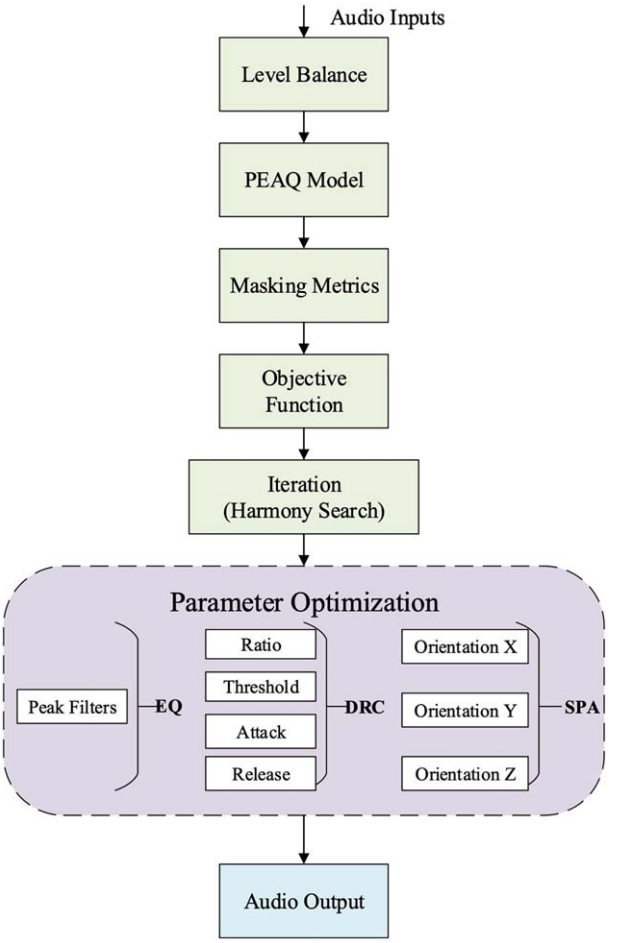


Fig. 1. The workflow of the proposed system. First, the level balance will adjust the loudness of each track. Next, the PEAQ model quantifies the masking metrics, which are subsequently incorporated into the objective function. This optimization process, driven by the Harmony Search algorithm, iteratively adjusts the parameters of the applied audio effects, including EQ, DRC, and SPA. The system continuously refines these parameters by reevaluating the computed masking values to optimize auditory clarity.

The following section outlines the workflow and components of the proposed method, designed to effectively address these challenges.

2 METHOD

Given multiple audio inputs (SEC. 2.1), the goal is to apply audio effects with optimized parameters to minimize the unwanted auditory masking. The workflow of the whole system[2] is shown in Fig. 1. The first step of the system is level balance (SEC. 2.2). After that, the PEAQ model (SEC. 2.3) and masking metrics (SEC. 2.4) will be used to measure auditory masking in the system. To minimize the masking, an objective function (SEC. 2.5) combined with harmony searching is used to (SEC. 2.6) optimize the parameters of

[2]https://github.com/xl2591/AutoMix.

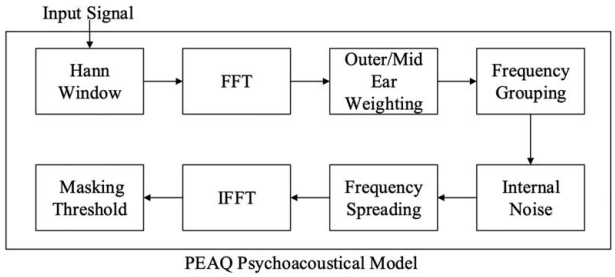


Fig. 2. The workflow of the PEAQ psychoacoustic model.

applied audio effects (Sec 2.7) including EQ, DRC, and SPA.

2.1 Audio Input

This study is applicable to multispeaker scenarios. In the following sections, a track refers to an individual speaker. In general speech scenarios, it serves as input for communication systems, whereas in this study, it specifically represents a recording of a single speaker. Detailed information about the experimental stimuli is provided in Sec 3.2.

2.2 Level Balance

In the work of [11], the authors compared different audio effects across multiple scenarios and found that achieving equal level balance has the most significant impact on reducing auditory masking in multispeaker environments. In achieving level balance, the present authors use Loudness Units Full-Scale (LUFS) as a standardized measure for evaluating sound loudness. LUFS is a measure that accounts for both human perception and electrical signal strength. According to the *EBU Recommendation 128* guidelines [37], the recommended loudness level for radio programs is set at –23 LUFS. In practice, the loudness of each track is calculated and adjusted using the method described by [38]. This process ensures a consistent and optimal loudness experience across various tracks, adhering to industry standards for broadcasting, music production, streaming, podcasts, and other forms of loudness management.

2.3 PEAQ Model

To quantify the masking effect and optimize audio quality, obtaining the masking threshold is crucial. According to Hu et al. [36], PEAQ [9] offers a more accurate representation of human auditory perception compared to other psychoacoustic models for estimating masking thresholds. The PEAQ model simulates aspects of human hearing to estimate thresholds for masking of audio signals. Fig. 2 describes the workflow of the PEAQ model in the proposed system.

- **Frequency domain transformation:** The input signal is transformed into the frequency domain using the fast Fourier transform with a Hann window.
- **Loudness correction:** The spectrum energy loudness is corrected using a weighting function to simulate the human ear's sensitivity curve, considering the outer/mid ear function. This adjustment accounts for how sound pressure level affects perceptual quality.
- **Frequency grouping:** The transformed audio signal is divided into 109 frequency bands according to the Bark scale, as defined in the basic version of the PEAQ model [9].
- **Internal noise simulation:** The internal noise component simulates the noise produced by blood flow in the inner ear, as described in [9].
- **Frequency spreading:** This step simulates the smearing effect of wide auditory filters, reflecting how auditory filters spread energy across frequencies.
- **Estimation of masking threshold:** Eq. (1) gives the masking threshold for the kth frequency band on the Bark scale [36].

$$E_{\text{mask}(k)} = E_{f(k)} - m_{(k)},$$
$$m_{(k)} = \begin{cases} 3 & z \le z_L + 12, \\ 0.25(z - z_L) & z > z_L + 12, \end{cases} \tag{1}$$

where $E_{\text{mask}(k)}$ is the masking threshold, $E_{f(k)}$ is the energy response, $m_{(k)}$ adjusts the amplitude of the kth frequency band, all in decibel units, z is the central frequency of each band (provided by PEAQ with 109 bands), converted to the Bark scale, and z_L is equal to 0.8594 [36].

All the above steps pertain to processing a single frame, while the time spreading part works on multiple frames. With these values computed, the final masker-to-signal ratio (MSR) in each frequency band (fb) is defined as, according to [34],

$$MSR(fb) = 10\log_{10}\left(\frac{E_{\text{mask}}(fb)}{E_f(fb)}\right). \tag{2}$$

2.4 Masking Metrics

In the multiple speaker scenario, to quantify the masking value for each track and to estimate how much a track is masked by other tracks, Parker and Fenton's approach [32] is used in Eq. (3).

$$T_n'(fb) = H\left(\sum_{\substack{i=1 \\ i\neq n}}^{N} S_i\right). \tag{3}$$

In this equation, N is the total number of tracks, S_i is the signal from the ith track, n is the track being masked by all other tracks, $T_n'(fb)$ represents the masking threshold of track n, which is influenced by the sum of accompanying tracks, and H represents all the mathematical operations in the MPEG psychoacoustic model used to calculate the masking threshold.

In Eq. (4), the $E_{f,n}(fb)$ is the energy of track n computed in each frequency band as described in Eq. (1). $T_n'(fb)$ replaces the $E_{\text{mask}}(fb)$ in Eq. (2). The final MSR in each

frequency band, how much track n is masked by other tracks in each frequency band is defined as

$$MSR_n(fb) = 10\log_{10}\left(\frac{T'_n(fb)}{E_{f,n}(fb)}\right). \quad (4)$$

The range maximum amount of masking distance value, T_{max}, is set to 20 dB [32]. Then M_n, the cross-adaptive multitrack masking measurement for track n, is given by

$$M_n = \sum_{fb \subset E_{f,n} < T'_n(fb)} \frac{MSR_n(fb)}{T_{max}}. \quad (5)$$

2.5 Objective Function

The aim of the objective function is to decrease auditory masking through parameter optimization. Each track's parameter should be computed until the masking value is reduced to minimum or the system reaches the maximum iterations' number. $\mathbf{x_C}$ is the function of all tracks' parameter control. The value of masking metrics is given by $M_i(\mathbf{x_C})$, which is the masking value for i tracks comparing with other sum tracks. The total amount of masking is $M_T(\mathbf{x_C})$, which consists of the sum of $M_i^2(\mathbf{x_C})$ for $i = 1$ to N,

$$M_T(\mathbf{x_C}) = \sum_{i=1}^{N} M_i^2(\mathbf{x_C}). \quad (6)$$

The objective of Eq. (6) is to minimize the sum of the masking across tracks and so can be used as the first part of the objective function. The second objective is that the masking is balanced. This means no difference between masking levels and a maximum masking difference is given by

$$M_d(\mathbf{x_C}) = \max_{i,j \in \{1,\ldots,N\}, i \neq j} (\left\| M_i(\mathbf{x_C}) - M_j(\mathbf{x_C}) \right\|). \quad (7)$$

From the parameter changes, the value of $\mathbf{x_C}$ will influence not only the masking effect on the track itself but also the masking effects on all other tracks. The optimal $\mathbf{x_C^*}$ is finally defined as

$$\mathbf{x_C^*} = \min_{\mathbf{x_C}}[M_T(\mathbf{x_C}) + M_d(\mathbf{x_C})]. \quad (8)$$

2.6 Iteration

In the iteration section, the Harmony Search optimization algorithm is used for this system. The Harmony Search algorithm is a music-inspired metaheuristic optimization method [10] that uses mutation and selection to gradually improve the existing solution by continuously adjusting and combining the candidate solutions until the optimal solution is found. The process of running the algorithm is similar to the harmonization process in a band performance: each musical instrument represents a decision variable, a musical note corresponds to a variable value, and a harmony represents a solution vector. By continuously coordinating and adjusting the performance of each musician, one hopes to find an optimal performance combination.

To enhance the convergence speed and computational efficiency of the algorithm, Harmony Search is integrated with integer optimization [39]. This integration facilitates an efficient exploration of the solution space, enabling the identification of the optimal solution. Recalling Eq. (8), the ith component of $\mathbf{x_c}$ resets during the iteration following Eq. (9):

$$\mathbf{x_C^*} = \left(\left\lfloor \frac{\mathrm{U} \cdot (\mathrm{maxV} - \mathrm{minV})}{\mathrm{step}} \right\rfloor + 1\right) \cdot \mathrm{step} + \mathrm{minV}. \quad (9)$$

In Eq. (9), the step determines the magnitude of change applied to each parameter, influencing the optimization convergence and the ability to fine-tune the effects. The maxV and minV is the maximum and minimum value in the value range of audio effect parameters (as shown in Table 1). U is a random variable uniformly distributed over the interval [0, 1].

Table 1. The value range of audio effects (EQ, DRC, and SPA) parameters.

Audio Effects	Min Value	Max Value	Step
EQ gain bands 1–8	–15 dB	15 dB	3
DRC ratio	1	5	1
DRC threshold	–15 dB	0 dB	3
DRC attack	0.01 s	0.5 s	0.001
DRC release	0.05 s	1 s	0.01
SPA x-axis	–3	3	0.5
SPA y-axis	–3	3	0.5
SPA z-axis	–3	3	0.5

2.7 Audio Effects

The Harmony Search algorithm will randomly select different parameters with different effects in EQ, DRC, and SPA through the Web Audio API [40]. In the EQ stage, each input signal undergoes gain modification using second-order infinite impulse response filters within a filter bank comprising eight frequency bands. The center frequencies of the equalizers were set at 60; 100; 200; 400; 800; 1,600; 2,500; and 7,500 Hz, covering the typical frequency range of human speech [41]. These frequency bands are approximately distributed in octave intervals, aligning with the long-term average speech spectrum [42]. This design ensures that key speech components are adequately captured and enhanced.

The SPA employs Cartesian coordinates (x, y, z-axis) for source positioning, using vectors for location and a 3D directional cone for orientation. To establish the parameter range, consultations were conducted, followed by iterative testing until optimal results were achieved. The audio engineers involved in this process included three professionals, two female and one male, with an average age of 28. The parameter ranges for each audio effect are presented in Table 1.

3 EVALUATION

Each scenario is edited, with its corresponding soundtracks redesigned and synchronized with the visuals. The soundtracks will serve as stimuli for the listening tests. Each scenario has a distinct number of tracks: three tracks for teleconferencing, four tracks for gaming, and six tracks

Table 2. The loudness results for the objective test 1. The LUFS comparison of the voice tracks before and after level balancing in teleconferencing scenarios.

File Name	Loudness Before	Loudness After
Total Track	–12.172	–14.940
Track1	–27.279	–19.882
Track2	–12.655	–21.751
Track3	–44.064	–20.343

for live streaming. These files have a sample rate of 48 kHz, are single-channel, and have a bit depth of 16 bits.

3.1 Objective Test 1

Three objective tests were conducted to analyze the results concerning loudness, long-term average spectrum, and spatial positions for various tracks. Due to page limitations, the analysis is illustrated using the stimulus from the teleconferencing scenario as an example, to compare the output from Unmix (unprocessed audio) and Automix (the proposed system).

3.1.1 Loudness Analysis

Table 2 presents the loudness levels of each track before and after automatic mixing. The system successfully balanced the loudness levels according to the LUFS standard, ensuring a consistent and well-balanced audio foundation for subsequent mixing processes. By adjusting the loudness levels, the system ensures that no single track dominates: each speaker is clearly audible without overshadowing others.

3.2 Stimuli

To simulate multiple speaker scenes, the audio stimulus was collected from the LibriSpeech dataset [43] or extracted from multiple-speakers video in YouTube. Three scenarios were designed: teleconferencing, gaming, and live streaming, as shown in Fig. 3.

3.2.1 Spatial Analysis

Fig. 4 illustrates the spatial locations of the tracks following the automixing process. Before automixing, the tracks were centered, resulting in overlapping sounds with similar frequency components. This overlap may cause interference or masking of certain frequencies, making them indistinct to the human ear. After the process, their spatial positions became more distinctly separated. This separation enhances the clarity and discernibility of each individual track within the audio scene.

3.2.2 Frequency Analysis

Fig. 5 shows that the "Unmix" figure indicates that Track 2 might dominate the frequency range from 150 to 10,000 Hz, potentially overshadowing other sounds. This dominance could lead to Track 2 becoming the primary focus, with other sounds being masked or subdued. In contrast, the "After Mix" plot demonstrates the effectiveness of this

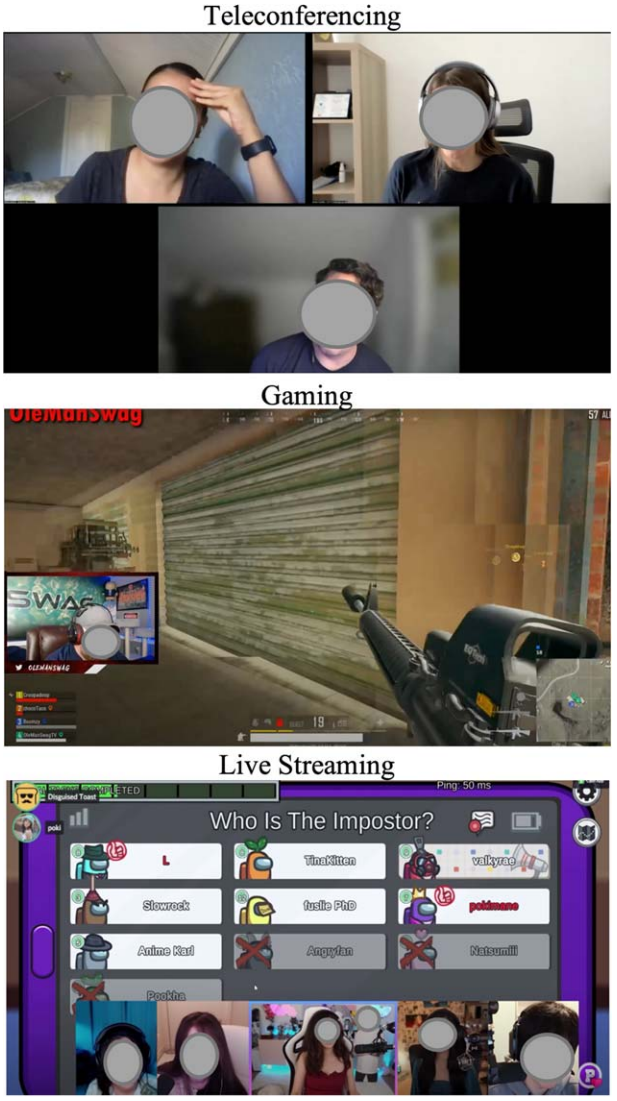


Fig. 3. Three scenarios of stimuli: three speakers in a teleconferencing scenario, four speakers in a gaming environment, and six speakers in a live-streaming debate setting.

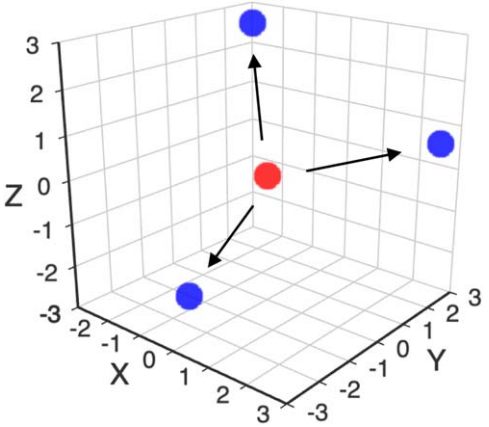


Fig. 4. The spatial results for the objective test 1. Red points represent the spatial locations before automixing, while blue points represent the locations after automixing.

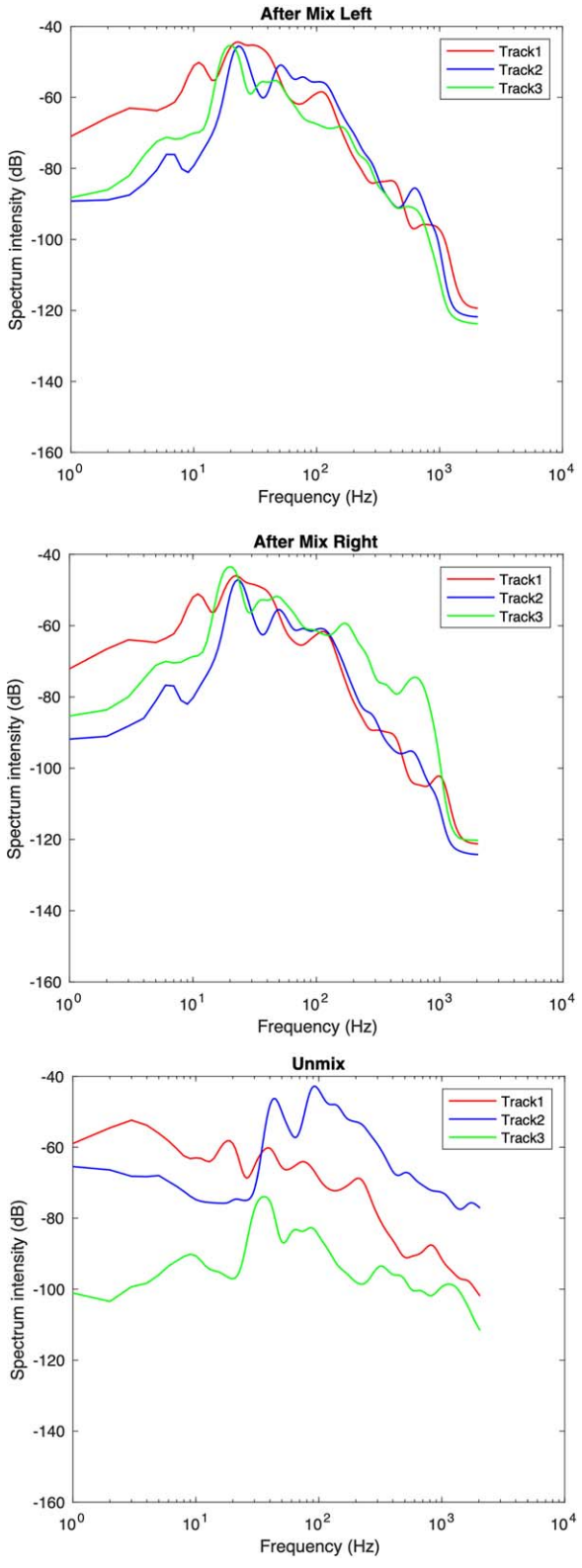


Fig. 5. The spectrum results for the objective test 1.

speech enhancement system, where the frequency distribution is well-balanced across the left and right channels.

The mixed system achieves a more cohesive spatial arrangement, ensuring that no single track overwhelms the others. This balance not only allows all elements of the audio to be clearly perceptible but also supports the system's goal of minimizing auditory masking. This system enables listeners to easily focus on or shift their attention between multiple simultaneous speakers, especially in complex scenarios such as teleconferencing, which often involves multiple speakers.

Table 3. The score of STI and MOSNet in different scenarios, higher value indicating better speech quality. The highest objective test value for each scene is highlighted in bold.

Scene	File Name	MOSNet	STI
Gaming	Automatic	2.92	0.51
	Manual	**3.01**	0.61
	Unmix	2.76	0.44
	Existing Mix	2.58	**0.73**
Teleconferencing	Automatic	**3.25**	**0.60**
	Manual	3.14	0.55
	Unmix	3.04	0.51
	Existing Mix	3.22	0.55
Live streaming	Automatic	3.01	**0.75**
	Manual	**3.04**	0.63
	Unmix	3.03	0.51
	Existing Mix	3.00	0.60

3.3 Objective Test 2

To further evaluate the effectiveness of the system, two speech metrics were considered: Speech Transmission Index (STI) [44] and Mean Opinion Score Network (MOSNet) [45] to assess the intelligibility of the results. The materials for the objective test included the output from the proposed automatic mixing system, the output from an existing automatic mixing system [11], a manual mix output by an expert audio engineer, and an unmixed version of the content.

The STI ranges from 0 to 1, with higher values indicating better transmission conditions. It reflects the potential intelligibility of speech conveyed through a system. Traditionally, STI is evaluated by transmitting specially modulated test signals through the system. Subsequently, the received output is analyzed to assess potential speech intelligibility. However, some systems are not suitable for playing or recording the test signals. To address this limitation, previous research [46–48] has proposed using speech signals as probe stimuli instead of these special modulated signals, mitigating certain constraints of the traditional STI approach with varying degrees of success. Based on this idea, this study directly utilizes speech stimuli in the STI model to assess speech clarity.

The MOSNet value ranges from 1 to 5, with higher scores representing better speech quality. The result of the test stimuli as shown in Table 3.

In Table 3, the objective evaluation results (STI and MOSNet) are presented under four mixing conditions for each scenario. "Automatic" refers to the proposed system, "Unmix" denotes the unprocessed input, "Existing Mix" represents the existing automatic mixing system's output from [11], and "Manual" corresponds to the human mix created by a professional engineer. The results show that:

- **Game scenario:** The Automatic mixing method achieved an STI score of 0.51, which, although lower than Manual (0.61) and Existing Mix (0.73), was notably higher than Unmix (0.44). This indicates that the Automatic method did not achieve high intelligibility in this context. The MOSNet score for Automatic mixing was 2.92, ranking second only to Manual (3.01), and considerably higher than Existing Mix (2.58). These results suggest that the Automatic method maintained a relatively high perceptual quality even though its intelligibility was limited in fast-paced, interaction-heavy scenarios like gaming.
- **Teleconferencing scenario:** The Automatic method excelled in this context, outperforming both the Manual and Unmix methods. It provided the best balance between audio quality and intelligibility, making it particularly effective for teleconferencing, where clear communication is essential.
- **Live-streaming scenario:** The Automatic mixing method again achieved the highest STI score of 0.75, clearly outperforming other systems. Although the MOSNet scores across all methods were close (around 3.00), the Automatic method demonstrated a clear advantage in intelligibility.

In both the teleconferencing and live-streaming scenarios, the Automatic method stated strong performance in terms of intelligibility. While the predicted audio quality in the live-streaming scenario was slightly lower than that of the Manual method, the high STI score highlights the Automatic system's strength in enhancing intelligibility. These results suggest that the Automatic system is often a reliable solution that can reduce the need for manual adjustments. In contrast, the Unmix method showed the weakest intelligibility performance across scenarios.

3.4 Subjective Test

In the subjective listening test, the Go Listen platform [49] was used to conduct a blind comparison test. The materials of the subjective listening test included the current automatic mixing system's output, the previous automatic mixing system's output from [11], manual mix output by an expert audio engineer, one from the unmixed content, and two hidden anchor versions of the unmixed content (3.5-kHz and 7-kHz low-pass filters) were used to calibrate participants' rating scales and enable indirect screening based on their ratings [50].

3.4.1 Tester

A total of 18 participants took part in the test and were instructed to conduct the evaluation in a quiet environment. Data from two participants were considered invalid and excluded from the analysis due to noncompliance with the experimental requirements. In total, the subjective listening test involved 16 participants, with nine males and seven females included after excluding noncompliant data. All participants had normal hearing, confirmed by a hearing test question that ensured no significant hearing loss. The participants had an audio or musical background ranging from 3 to 15 years.

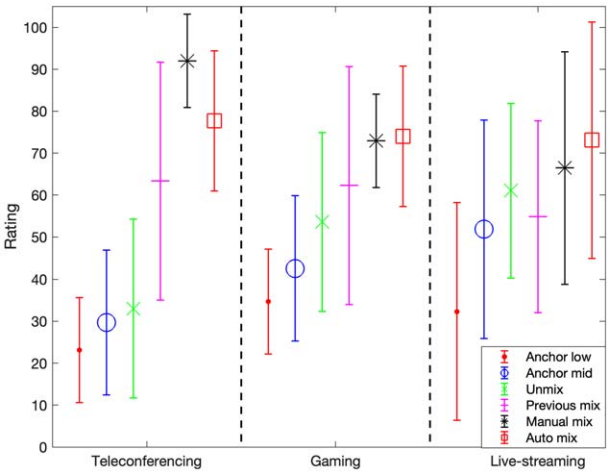


Fig. 6. The multistimulus test results, with 95% confidence intervals.

Table 4. ANOVA results for different scenarios. Bold values indicate statistical significance ($p < 0.05$).

Scene	F Value	p Value
Teleconferencing	37.01	<**0.001**
Gaming	7.93	<**0.001**
Live streaming	5.00	<**0.001**

The test was conducted in a quiet, soundproof room to minimize environmental interference, and all participants used monitor headphones, such as the AKG K702, during the test. During the listening test, participants were informed about the specific scenario, including the number of speakers. They were then asked to assess audio clarity by determining whether they could clearly perceive and distinguish each speaker within the given context.

Results are presented in Fig. 6. The Automix has consistent ratings around 75 and outperforms the Unmix, Anchors, and Existing Automatic system in each scenario. In the teleconferencing scenario, the manual mix gets the highest average score. However, the average score of the manual mix is below the Automix in both gaming and live-streaming scenarios. A potential reason for this is the added complexity when attempting to manually process a larger number of tracks.

The Kolmogorov-Smirnov test [51] indicates that the current data follow a normal distribution. For the next step of analysis, a one-way analysis of variance (ANOVA) was conducted to analyze the variations among the files within three groups of experimental scenes. Table 4 indicated statistically significant differences within each group.

To further analyze the results, Tukey's Honestly Significant Difference post hoc test was applied to calculate the p values for pairwise comparisons, specifically assessing the significant differences between files and the proposed system (Automix) across three scenarios. As shown in Table 5,

Table 5. Post hoc analysis of mean differences and p values between files and the proposed system across three scenarios. Bold values indicate statistical significance ($p < 0.05$).

	Teleconferencing		Gaming		Live Streaming	
File Name	Mean Difference	p Value	Mean Difference	p Value	Mean Difference	p Value
Anchor low	–54.5625	<**0.001**	–39.3125	<**0.001**	–40.8125	<**0.001**
Anchor mid	–48.0000	<**0.001**	–31.4375	<**0.001**	–21.2500	0.2263
Unmix	–44.6875	<**0.001**	–20.3750	0.2099	–12.0625	0.8106
Existing mix	–14.3065	0.2683	–11.6875	0.6587	–18.2500	0.3081
Manual mix	14.3125	0.2683	–1.0625	1.0000	–3.3750	0.9999

the average score difference was included to highlight the variations.

As expected, Anchor Low and Anchor Mid received significantly lower scores compared to the tested systems. Moreover, the proposed system consistently outperformed the Existing Mix across all three scenarios and achieved performance closely approaching that of the Manual mix. The Unmix system demonstrated consistent underperformance relative to the proposed system, particularly in the teleconferencing and gaming scenarios.

3.4.2 ANOVA Significance Analysis

Notably, in the livestreaming scenario, a significant difference was observed in only one group. This reveals that the results are affected by the stimuli scenarios. This can be attributed to differences in the number of speakers across the scenarios: six speakers in the livestreaming scenario, four in the gaming scenario, and three in the teleconferencing scenario. It can be inferred that participants' comprehension declined accordingly with the increase in the number of speakers.

Overall, this Automix system has shown good adaptability across these application scenarios, outperforming other automatic mixing technologies. Although Automix has surpassed manual mixing in gaming and live-streaming scenarios, there remains room for optimization to achieve or exceed the performance of Manual mixing in all scenarios. Future improvements might focus on better handling of complex audio environments, such as those involving a large number of tracks, to further improve the overall performance of the system.

4 CONCLUSION

The authors have developed a novel automatic speech enhancement system tailored for scenarios involving multiple speakers. This system is capable of presenting multiple tracks of speech information, allowing users to focus on a specific track while seamlessly shifting their attention to another as needed. This system integrates three key aspects: spatial, frequency, and loudness processing to minimize auditory masking between tracks. Building on previous research and adhering to industry standards for broadcasting, this system can automatically adjust the loudness of each track. Additionally, by leveraging adaptive masking metrics derived from PEAQ, the system applies three audio effects: EQ, DRC, and SPA, along with Harmony Search and integer optimization, to optimize parameter settings.

Objective tests were conducted using two speech metrics: MOSNet and STI. Furthermore, 16 professional audio experts participated in subjective evaluations, comparing the audio clarity produced by this system against existing automix solutions and manual mixing in multispeaker scenarios. The results demonstrate that this system is competitive with both the existing solution and manual mixing. This study provides valuable insights into a lightweight automatic multispeaker mixing system. For future work, it is essential to enhance the algorithm's robustness for real-time implementation and to develop effective strategies for adapting parameter adjustments to dynamic changes in the audio environment.

5 ACKNOWLEDGMENT

Sincere gratitude is extended to the three audio engineers, Dr. Angeliki Mourgela, Fangting Li, and DengDeng Bao, as well as to Dr. Shuren Tan and Dr. Yin-Jyun Luo for their valuable feedback on this paper.

6 REFERENCES

[1] B. Arons, *A Review of the Cocktail Party Effect*, vol. 12 (MIT Press, Cambridge, MA, 1992).

[2] E. C. Cherry, "Some Experiments on the Recognition of Speech, With One and With Two Ears," *J. Acoust. Soc. Am.*, vol. 25, no. 5, pp. 975–979 (1953 Sep.). https://doi.org/10.1121/1.1907229.

[3] Y. C. Chen, L. Wang, and S. Tsai, "Co-Talker Separation Using the 'Cocktail Party Effect,'" *J. Audio Eng. Soc.*, vol. 44, no. 12, pp. 1084–1096 (1996 Dec.).

[4] B. Wiem, B. M. M. Anouar, and B. Aicha, "Hybrid Approach to Speech Source Separation Depending on the Voicing State," *J. Audio Eng. Soc.*, vol. 66, no. 12, pp. 1041–1050 (2018 Dec.). https://doi.org/10.17743/jaes.2018.0059.

[5] S. Araki, T. Hayashi, M. Delcroix, et al., "Exploring Multi-Channel Features for Denoising-Autoencoder-Based Speech Enhancement," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 116–120 (South Brisbane, Australia) (2015 Apr.). https://doi.org/10.1109/ICASSP.2015.7177943.

[6] R. Y. Litovsky, M. J. Goupell, S. M. Misurelli, and A. Kan, *Hearing With Cochlear Implants and Hearing Aids in Complex Auditory Scenes* (Springer, Cham, Switzerland, 2017).

[7] G. Courtois, P. Marmaroli, M. Lindberg, Y. Oesch, and W. Balande, "Implementation of a Binaural Localization Algorithm in Hearing Aids: Specifications and Achievable Solutions," presented at the *136th Convention of the Audio Engineering Society* (2014 Apr.), paper 9034.

[8] M. S. Kavalekalam, M. G. Christensen, F. Gran, and J. B. Boldt, "Kalman Filter for Speech Enhancement in Cocktail Party Scenarios Using a Codebook-Based Approach," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 191–195 (Shanghai, China) (2016 Mar.). https://doi.org/10.1109/ICASSP.2016.7471663.

[9] T. Thiede, W. C. Treurniet, R. Bitto, et al., "PEAQ—The ITU Standard for Objective Measurement of Perceived Audio Quality," *J. Audio Eng. Soc.*, vol. 48, no. 1/2, pp. 3–29 (2000 Jan.).

[10] X. S. Yang, *Nature-Inspired Optimization Algorithms* (Elsevier, Oxford, UK, 2014).

[11] L. Xiaojing, J. Reiss, and A. Mourgela, "An Automatic Mixing System for Teleconferencing," presented at the *153rd Convention of the Audio Engineering Society* (2022 Oct.), paper 24.

[12] A. Griffin, T. Hirvonen, C. Tzagkarakis, A. Mouchtaris, and P. Tsakalides, "Single-Channel and Multi-Channel Sinusoidal Audio Coding Using Compressed Sensing," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 19, no. 5, pp. 1382–1395 (2011 Jul.). https://doi.org/10.1109/TASL.2010.2090656.

[13] B. Tolooshams, R. Giri, A. H. Song, U. Isik, and A. Krishnaswamy, "Channel-Attention Dense U-Net for Multichannel Speech Enhancement," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 836–840 (Barcelona, Spain) (2020 Apr.). https://doi.org/10.1109/ICASSP40776.2020.9053989.

[14] C. Fan, H. Zhang, A. Li, et al., "CompNet: Complementary Network for Single-Channel Speech Enhancement," *Neural Netw.*, vol. 168, pp. 508–517 (2023 Nov.). https://doi.org/10.1016/j.neunet.2023.09.041.

[15] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "Eigenvector-Based Speech Mask Estimation for Multi-Channel Speech Enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 12, pp. 2162–2172 (2019 Dec.). https://doi.org/10.1109/TASLP.2019.2941592.

[16] R. J. M. van Hoesel and G. M. Clark, "Evaluation of a Portable Two-Microphone Adaptive Beamforming Speech Processor With Cochlear Implant Patients," *J. Acoust. Soc. Am.*, vol. 97, no. 4, pp. 2498–2503 (1995 Apr.). https://doi.org/10.1121/1.411970.

[17] Y. Geng, T. Zhang, M. S. Yaw, and H. Wang, "A Speech Enhancement Method Based on the Combination of Microphone Array and Parabolic Reflector," *J. Audio Eng. Soc.*, vol. 70, no. 1/2, pp. 5–23 (2022 Feb.). https://doi.org/10.17743/jaes.2021.0047.

[18] 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, no. 1, pp. 1–57 (2005 Jan.).

[19] H. Erdogan, J. R. Hershey, S. Watanabe, M. I. Mandel, and J. Le Roux, "Improved MVDR Beamforming Using Single-Channel Mask Prediction Networks." in *Proceedings of INTERSPEECH*, pp. 1981–1985 (San Francisco, CA) (2016 Sep.).

[20] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural Network Based Spectral Mask Estimation for Acoustic Beamforming," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 196–200 (Shanghai, China) (2016 May). https://doi.org/10.1109/ICASSP.2016.7471664.

[21] , C. R. Mason, T. L. Rohtla, and P. S. Deliwal, "Release From Masking Due to Spatial Separation of Sources in the Identification of Nonspeech Auditory Patterns," *J. Acoust. Soc. Am.*, vol. 104, pp. 422–431 (1998 Jul.). https://doi.org/10.1121/1.423246.

[22] J. Skowronek and A. Raake, "Assessment of Cognitive Load, Speech Communication Quality and Quality of Experience for Spatial and Non-Spatial Audio-Conferencing Calls," *Speech Commun.*, vol. 66, pp. 154–175 (2015 Feb.). https://doi.org/10.1016/j.specom.2014.10.003.

[23] A. Roginska and P. Geluso, *Immersive Sound: The Art and Science of Binaural and Multi-Channel Audio* (Routledge, New York, NY, 2018), 1st ed.

[24] M. Rothbucher, M. Kaufmann, J. Feldmaier, et al., "3D Audio Conference System With Backward Compatible Conference Server Using HRTF Synthesis," *J. Multim. Process. Technol.*, vol. 2, no. 4, pp. 159–175 (2011 Dec.).

[25] Z.-Q. Wang and D. Wang, "All-Neural Multi-Channel Speech Enhancement," in *Proceedings of INTERSPEECH*, pp. 3234–3238 (Hyderabad, India) (2018 Sep.). https://doi.org/10.21437/Interspeech.2018-1664.

[26] J. Heymann, L. Drude, and R. Haeb-Umbach, "Neural Network Based Spectral Mask Estimation for Acoustic Beamforming," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 196–200 (Shanghai, China) (2016 May). https://doi.org/10.1109/ICASSP.2016.7471664.

[27] N. Ito, S. Araki, and T. Nakatani, "Permutation-Free Convolutive Blind Source Separation via Full-Band Clustering Based on Frequency-Independent Source Presence Priors," in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, pp. 3238–3242 (Vancouver, Canada) (2013 Oct.). https://doi.org/10.1109/ICASSP.2013.6638256.

[28] Y. Jiang, H. Zhou, and Z. Feng, "Performance Analysis of Ideal Binary Masks in Speech Enhancement," in *Proceedings of 4th International Congress on Image and Signal Processing*, vol. 5, pp. 2422–2425 (Shanghai, China) (2011 Apr.). https://doi.org/10.1109/CISP.2011.6100732.

[29] L. Pfeifenberger, M. Zöhrer, and F. Pernkopf, "Eigenvector-Based Speech Mask Estimation for Multi-Channel Speech Enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*,

vol. 27, no. 12, pp. 2162–2172 (2019 Dec.). https://doi.org/10.1109/TASLP.2019.2941592.

[30] F. Bao and W. H. Abdulla, "A New Ratio Mask Representation for CASA-Based Speech Enhancement," *IEEE/ACM Trans. Audio Speech Lang. Process.*, vol. 27, no. 1, pp. 7–19 (2019 Sep.). https://doi.org/10.1109/TASLP.2018.2868407.

[31] S. Srinivasan, N. Roman, and D. Wang, "Binary and Ratio Time-Frequency Masks for Robust Speech Recognition," *Speech Commun.*, vol. 48, no. 11, pp. 1486–1501 (2006 Nov.).

[32] A. Parker and S. Fenton, "Musical Mix Clarity Prediction Using Decomposition and Perceptual Masking Thresholds," *Appl. Sci.*, vol. 11, no. 20, paper 9578 (2021 Oct.). https://doi.org/10.3390/app11209578.

[33] K. Kumar, R. Pandey, M. Karthik, S. S. Bhattacharjee, and N. V. George, "Robust and Sparsity-Aware Adaptive Filters: A Review," *Signal Process.*, vol. 189, paper 108276 (2021 Dec.). https://doi.org/10.1016/j.sigpro.2021.108276.

[34] Z. Ma, *Intelligent Tools for Multi-Track Frequency and Dynamics Processing*, Ph.D. thesis, Queen Mary University of London, London, UK (2016 Jun.). https://qmro.qmul.ac.uk/xmlui/handle/123456789/23289.

[35] D. J. LeGall, "MPEG: A Video Compression Standard for Multimedia Applications," Commun. ACM, vol. 34, no. 4, pp. 46–58 (1991 Apr.). https://doi.org/10.1145/103085.103090.

[36] X. Hu, G. He, and X. Zhou, "PEAQ-Based Psychoacoustic Model for Perceptual Audio Coder," in *Proceedings of the International Conference on Advanced Communications Technology (ICACT)*, pp. 1594–1598 (Phoenix Park, South Korea) (2006 Feb.). https://doi.org/10.1109/ICACT.2006.206344.

[37] Union European Broadcasting, "Loudness Normalisation and Permitted Maximum Level of Audio Signals," *EBU Recommendation 128* (2020 Aug.).

[38] C. J. Steinmetz and J. Reiss, "Pyloudnorm: A Simple Yet Flexible Loudness Meter in Python," presented at the *150th Convention of the Audio Engineering Society* (2021 May), paper 10483.

[39] G. Sierksma and Y. Zwols, *Linear and Integer Optimization: Theory and Practice* (CRC Press, Boca Raton, FL, 2015).

[40] J. D. Reiss, *Working With the Web Audio API* (Routledge, Abingdon, UK, 2022).

[41] E. Jacewicz, J. M. Alexander, and R. A. Fox, "Introduction to the Special Issue on Perception and Production of Sounds in the High-Frequency Range of Human Speech," *J. Acoust. Soc. Am.*, vol. 154, no. 5, pp. 3168–3172 (2023 Nov.). https://doi.org/10.1121/10.0022496.

[42] D. Byrne, H. Dillon, K. Tran, et al., "An International Comparison of Long-Term Average Speech Spectra," *J. Acoust. Soc. Am.*, vol. 96, no. 4, pp. 2108–2120 (1994 Oct.). https://doi.org/10.1121/1.410152.

[43] H. Zen, V. Dang, R. Clark, et al., "LibriTTS: A Corpus Derived From LibriSpeech for Text-to-Speech," *arXiv preprint arXiv:1904.02882* (2019 Apr.).

[44] K. L. Payton and E. C. Braida, "A Method to Determine the Speech Transmission Index From Speech Waveforms," *J. Acoust. Soc. Am.*, vol. 106, no. 6, pp. 3637–3648 (1999 Dec.). https://doi.org/10.1121/1.428216.

[45] C.-C. Lo, T.-Y. Hsiao, H. Kawai, and J.-H. Chou, "MOSNet: Deep Learning Based Objective Assessment for Voice Conversion," *arXiv preprint arXiv:1904.08352* (2019 Apr.).

[46] C. Ludvigsen, C. Elberling, G. Keidser, and T. Poulsen, "Prediction of Intelligibility of Non-Linearly Processed Speech," *Acta Oto-Laryngol.*, vol. 109, no. sup469, pp. 190–195 (1990 Jan.). https://doi.org/10.1080/00016489.1990.12088428.

[47] V. Hohmann and B. Kollmeier, "The Effect of Multichannel Dynamic Compression on Speech Intelligibility," *J. Acoust. Soc. Am.*, vol. 97, no. 2, pp. 1191–1195 (1995 Feb.). https://doi.org/10.1121/1.413092.

[48] L. Yang, J. Zhang, and Y. Yan, "An Improved STI Method for Evaluating Mandarin Speech Intelligibility," in *Proceedings of the International Conference on Audio, Language and Image Processing*, vol. 154, pp. 102–106 (Shanghai, China) (2008 Jul.). https://doi.org/10.1109/ICALIP.2008.4590080.

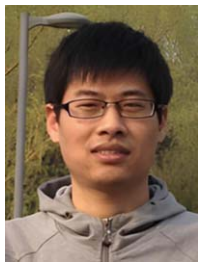
[49] D. Barry, Q. Zhang, P. W. Sun, and A. Hines, "Go Listen: An End-To-End Online Listening Test Platform," *J. Open Res. Softw.*, vol. 9, no. 1, paper 20 (2021 Jan.). https://doi.org/10.5334/jors.361.

[50] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," *Recommendation ITU-R BS.1534-3* (2015 Oct.).

[51] H. W. Lilliefors, "On the Kolmogorov-Smirnov Test for Normality With Mean and Variance Unknown," *J. Am. Stat. Assoc.*, vol. 62, no. 318, pp. 399–402 (1967 Apr.). https://doi.org/10.1080/01621459.1967.10482916.

THE AUTHORS


Xiaojing Liu


Hongwei Ai


Joshua D. Reiss

Xiaojing Liu is a researcher, music performer, and audio technologist. She has published two conference papers: "An Automatic Mixing System for Teleconferencing" at the 154th AES Convention and "User Preference Evaluation of the Masking Ratio in Multiple Speaker Scenarios" at the 156th AES Convention, funded by the QMUL Postgraduate Research Fund. Additionally, she serves as a conference reviewer for ACM CHI and ICME and as a Laboratory Demonstrator for the courses ECS661U User Experience Design and ECS602U/ECS707P Foundation of Digital Signal Processing at Queen Mary University of London. Currently, she is conducting Ph.D. research focusing on automatic mixing for gaming, teleconferencing, and streaming.

•

Hongwei Ai holds a master's degree in Computer Science from Peking University. His research focuses on the areas of artificial intelligence and machine learning, exploring innovative solutions to complex computational challenges. He is particularly interested in developing algorithms and models that improve machine learning efficiency and adaptability across various applications.

•

Joshua D. Reiss is Professor of Audio Engineering with the Centre for Digital Music at Queen Mary University of London. He has published more than 200 scientific papers (including over 50 in premier journals and seven best paper awards) and coauthored three books. His research has been featured in dozens of original articles and interviews on TV, on the radio, and in the press. He is a Fellow and Past President of the Audio Engineering Society and formal chair of their Publications Policy Committee. He cofounded the highly successful spin-out company, LandR, and recently cofounded Waveshaper AI, RoEx, and Nemisindo, also based on his team's research. He maintains a popular blog, YouTube channel, and Twitter feed for scientific education and dissemination of research activities.