Audio Engineering Society

Convention Express Paper 24

Presented at the 153rd
Convention 2022 October

An automatic mixing system for teleconferencing

Xiaojing Liu, Joshua D. Reiss and Angeliki Mourgela

Centre for Digital Music, Queen Mary University of London, London, UK

Correspondence should be addressed to Xiaojing Liu and Joshua D. Reiss (xiaojing.liu@qmul.ac.uk, joshua.reiss@qmul.ac.uk)

ABSTRACT

This paper describes an automatic mixing system for improving audio quality in teleconferencing applications. The work was focused on applying audio effects such as multitrack level balancing, spatialization, and equalization in order to reduce speech masking, thus allowing simultaneous speakers to be heard in a teleconference. The system used the ITU-R BS.1770 loudness measurement method and cross-adaptive audio effects to achieve average level balancing. A novel Force-directed model was implemented to automatically set the virtual position of each source. The equalization method was based on spectral decomposition techniques and a target of equal average perceptual loudness in each frequency band. Subjective evaluation was performed in the form of a multi-stimulus listening test, which indicated that the proposed automatic mixing system could compete with a manual mix by an experienced sound engineer.

1 Introduction

Audio enhancement for teleconferencing applications usually focuses on automatic gain control technology, acoustic echo cancellation, and noise suppression for real-time communications [1]. However, in multiple-contributor communication scenarios, it is difficult to hear others' voices when there is more than one person speaking simultaneously. Existing audio systems may perform badly in such situations and may cause, confusion for the audience.

Auditory masking is a phenomenon whereby the perception of one sound is affected by the presence of another sound [2]. Such masking typically happen when two sounds occur within the same critical bandwidth [3][4]. The audibility of a source (maskee) will be affected if another source (masker) has a higher amplitude within a critical band than that of the maskee [5]. Such masking can be categorized into a variety of types. The most referred to form is simultaneous masking, in which the maskee sound occurs at the same time as the masking sound. In contrast, in forward masking or post-masking, a sound will be masked by a preceding sound. The maximum interval time for this phenomenon is 10ms. In backward masking or pre-masking, the sound is masked by a masking sound that follows the maskee (maximum interval time around 30ms) [1].

Moreover, a high frequency produces a greater masking effect than a low frequency [6]. In a multi-participant teleconferencing situation, frequency masking will cause the audio quality to become muddy due to the listener needing to parse a diversity of contents occupying similar spectral positions.

Unmasking in music

In music production, there are many research methods addressing masking issues. Previous work considered the effectiveness of equalization, stereo panning, and compression for unmasking with musical content [7]. However, they compared the audio effects individually and did not consider the whole system for unmasking work. Furthermore, a mixing solution that reduces masking in teleconferencing may require a quite different use of audio effects than a mix aimed at reducing masking with multitrack music content. Automatic mixing and intelligent tools have been suggested to address masking issues for almost every aspect of music production [8]. Automatic mixing was developed around 15 years ago as a means of reducing the burden of manual mixing work for musicians or audio engineers [9]. Compared with manual audio engineering, which requires extensive effort and experience to manually set appropriate parameters and apply audio effects, intelligent systems typically aim to extract dynamic features from an audio source, and use that information to determine the parameters of audio effects to apply. After that, the intelligent system will process the input signal in a manner similar to the actions performed by a trained engineer. However, in a multitrack mixing context, the characteristics of one track will affect other tracks, and single-track audio processing cannot solve this problem.

The work presented in [10] was focused on the use of an adaptive filter to adjust the levels of multiple sources in a mix, in order to reduce the masking of one source. But this would typically increase the masking on other sources.

Tom et al. [11] suggested an automatic system to minimize masking in a multitrack recording through spatialization. They used masking detection through spectral overlap and divided the audio frequency into low, mid and high bands. They then arranged the different frequency bands to set them in different location. Pestana took an extreme approach [12], devising a tool that positioned individual frequency bands from all sources across the stereo panning range.

Hafezi [13] found that automatic equalization can resolve masking in multitrack music production. Wakefield and Dewey [7] argued that panning is the preferable method for the unmasking of stereo mixing, but they also suggest that the method may not be appropriate if the lead vocal track is the key track and call for further investigation and experimentation in future work. In [14], Matz et al implemented and evaluated combinations of automatic mixing tools, as well as a harmonic exciter [15], to generate a mix. In their subjective tests, participants indicated that the automatic mixing tools could improve the sound balance and transparency of a mix.

Unmasking in speech

While previous research provides extensive knowledge and experience for the task of unmasking, it focused on music or instrument mixing, rather than voice unmasking.

For speech, there is considerably less research in solving masking in multiple microphone systems. Dugan [16] proposed using an automatic microphone system that can consider all the microphone input levels to eliminate masking of microphones and achieve voice enhancement through dynamic range compression. When people use a microphone, system gain is concentrated at that microphone and others are attenuated. Dugan argued that an automatic system can provide an instant reaction for every microphone, thus helping to avoid system reaction time, technical errors and cost of manual operation, and defects usually associated with voice-operated systems. However, Dugan’s approach removes all sources except the dominant speaker. Thus, it removes masking on just one source, by preventing any other source from being heard.

Quan Wang et al. [17] discussed a system which used a deep neural network to train a voice filter based on spectrum masking. The output of the network minimized the difference between the masked amplitude spectrogram and the target magnitude spectrogram. However, this and many other related studies focus on time-frequency masking, which is related but not the same as auditory masking. In contract, Zhou et al [18] performed single-channel speech enhancement based more directly on psychoacoustic masking.

Others have incorporated visual information. The research of Gu et al. [19] is based on multi-channel separation, visual modality for target, direction information, speaker characteristic, and spatialized audio with room impulse response. Wu et al. [20] presented work in multi-modal speech separation using TasNet (time-domain speech separation network) and extracted lip embedding from video stream.

All of these studies are focused on separating out the speech from one mixture/one environment/one audio stream for the purpose of unmasking, but do not consider multitrack scenarios.

Rothbucher, et al. used HRTF synthesis to place every speaker in a 3D space in the open-source VoIP conferencing software Mumble [21]. Their work resulted in a plugin which let users manually adjust their virtual location in a teleconferencing situation. In contrast, our system will automatically send users to different locations based on analysis of the spectral content from each speaker.

Proposed approach

The proposed system addresses masking in speech teleconferencing. It will focus on remote teleconferencing situations involving participants with individual computers and playback devices, rather than meetings with multiple participants in a sound-equipped meeting room.

The level balancing methods use the ITU-R BS.1770 loudness measurement method [22] and cross-adaptive audio effects to achieve average level balancing for each voice track, in order to reduce auditory masking due to level differences. The spatialization method allocates Azimuth and Elevation parameters to localize each source, based on the smoothed average fundamental frequency of each voice track, considering the number of user/tracks, the action of users joining or leaving, and long-term changes in spectral content. The equalization method is based on spectral decomposition techniques, and a target of equal average perceptual loudness in each frequency band, to avoid spectral masking.

This work has been implemented in the Web browser for real-time applications. Audio analysis and audio effects use the Web Audio API [23]. Testing is performed using input data from the LibriSpeech dataset [24], which consists of approximately 1000 hours of 16kHz English speech.

The latency of the system is analysed by considering the overlap length and the window size in the system (network latency was not yet measured). Subjective and objective tests compare the proposed automatic mix against original content, existing automatic mixing systems, and a manual mix. The proposed system will adapt to process any number of mono or stereo tracks, and at any sample rate.

To the authors' knowledge, this represents the first study of a whole automatic mixing system targeted at teleconferencing applications.

2 Algorithm

The automatic system passes the audio tracks through audio analysis and effect blocks and the tracks are processed separately. The whole system is composed of three main parts: level balance, equalization balance, and spatialization balance which are shown in Figure 1.

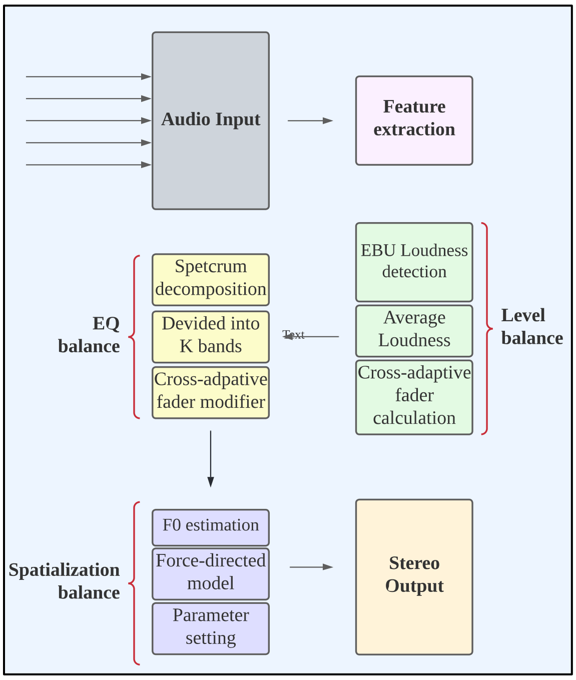


Figure 1: Block diagram of the proposed algorithm.

2.1Level Balance

The first step in the automatic mixing system is to average all the tracks' audio levels.

Loudness detection

According to ITU-R BS.1770 recommendation [22], LUFS stands for relative full-scale loudness units or full-scale loudness units (the maximum level the system can handle). It is a standardized measure of sound loudness that considers human perception and electrical signal strength. In the first, we use the k -weighting filter which consists of two bi-quadratic IIR filters, one is a high-shelf filter with corner frequency f_c = 1681hz and the other one is a high pass filter f_c = 38hz, to provide a frequency weighting for the psycho-acoustic model.

Fader Calculation

The total loudness of all tracks is summed and divided by the track count, to produce a dynamic target loudness that allows for intended fluctuations in the overall mix signal level. The gain applied to each track at sample n is given in Equation 1, where $G_m[n]$ is the gain applied to track m, $L_{av}[n]$ is the target loudness level, and $L_m[n]$ is the loudness level of track m.

$$G_m[n] = 10^{\frac{L_{av}[n]-L_m[n]}{20}} \qquad (1)$$

$$\text{where } L_{av}[n] = \sum_{m=1}^{M} L_m[n]/M$$

A set of three tracks, containing different levels of speech, were collected from the LibriSpeech dataset, and processed in real-time with our system. Fader values are then calculated as a ratio of the track loudness to the average loudness. As with other variables, the fader values are smoothed using an exponential moving average filter. Table 1 gives the measured loudness, in LUFS, of each track before and after level balancing. The loudness of each track has converged, resulting in a smaller standard deviation.

	Before level balancing	After level balancing
Track1	-33.8	-19.8
Track2	-14.9	-19.2
Track3	-50.5	-22.0
Standard deviation	**14.6**	**1.2**

Table. 1: The LUFS of the voice tracks before and after level balancing.

2.2Equalization.

We used a graphic equalizer to maintain equal average perceptual loudness within each frequency band, as in [25]. The equalization (EQ) method was based on spectral decomposition techniques. The first step is to decompose the spectrum of inputs through a filter bank consisting of 6 frequency bands: 20-60Hz, 60-200Hz, 600Hz-3kHz, 3-8kHz, and 8kHz and above.

The loudness of each frequency band was then estimated using LUFS, unlike [25] which used its own loudness estimation technique.

To test the effectiveness of the multitrack EQ and level balancing, we compared the magnitude spectra of input and output tracks. This is shown in Figure 2, which shows more similar spectra for the equalised tracks. Note that this could be problematic, since it could mean more overlapping content in the frequency domain, potentially leading to more masking.

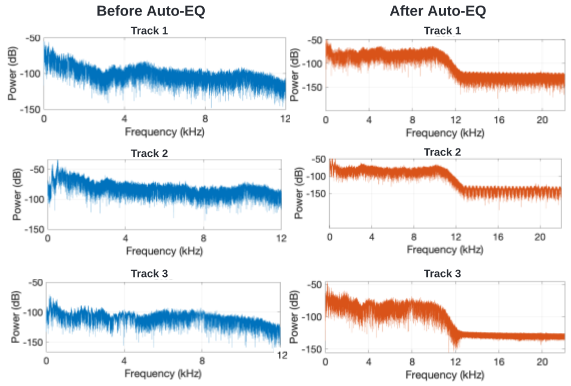


Figure. 2: Self-equalization of voice signals

2.3Localization methods

The spatialization method allocates Azimuth and Elevation parameters to localize each source, based on the smoothed average fundamental frequency of each voice track, considering the number of user/tracks, the action of users joining or leaving, and long-term changes in spectral content.

We estimated the fundamental frequency (F0) of each track using Yin's algorithm [26]. We then found differences in average F0 estimates for each pair of tracks in a multitrack, as shown in Table 2, which is used for spatialisation of each track.

	Track1	Track2	Track3
Track1	-	128.8	87.56
Track2	128.8	-	110.22
Track3	87.56	110.22	-

Table. 2: The difference, in Hertz, in fundamental frequencies for each pair of tracks for three speech tracks in a multitrack with simultaneous speakers.

The automatic localization method will position the sources using a force-directed layout model that considers the number of user/tracks in the system. Figure 3 depicts the resultant positions of tracks for multi-tracks with 3 voices and with 9 voices.

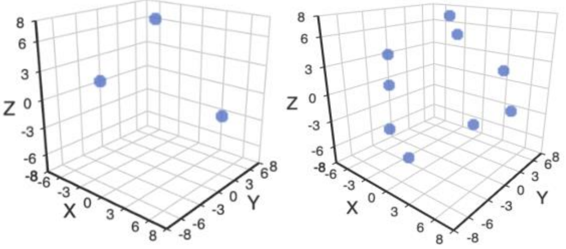


Figure. 3: Spatial positioning of sources for 3 tracks and for 9 tracks. The listener is placed at the origin.

Force-directed layout Model

We allocate the position of tracks on a sphere. The positions of sources in spatial audio may result in the loudness being unbalanced. Furthermore, the similarity of each track may result in masking. To address these issues, we chose the force-directed layout, also known as a spring embedder, which is commonly used to position sources on a graph surface [26]. This model is based on Coulomb's law.

$$F = k\frac{q_u q_v}{r^2} \qquad (2)$$

which gives the force F for two charges qu and qv separated by a distance r. k is the Coulomb constant.

In our context, $q_u q_v$ is replaced by the reciprocal of the difference in fundamental frequencies of the two tracks.

$$q_u q_v = \frac{1}{|\mathrm{F0_u} - \mathrm{F0_v}|} \qquad (3)$$

If $q_u q_v$ is large, the two tracks have close fundamental frequencies, so the distance between them should become large to avoid masking.

To apply the force-directed layout as in [26], the system will generate random positions for sources on a sphere around the listener. The geometric (Euclidean) 3D distance, between any two tracks will also be calculated. Positions are then updated iteratively by applying the force for all pairs of sources and in all directions.

Considering the intended use of the system in speech, our algorithm was further adapted to incremental force directed layout. If a new audio track is added to the system, the system will recalculate and resend the positions. However, changing all tracks' positions at the same time will result in a burst of noise at the beginning. We also allowed the algorithm to be adaptive, so that one may control the size of the sphere (i.e, room size), or vary the Coulomb constant k to change the strength of dispersion of sources around the sphere.

3 Implementation

All the test audio samples were extracted from the LibriSpeech dataset [23], which is derived from audiobooks and consists of approximately 1000 hours of English speech at 16kHz. We designed the system to run on the Web as shown in Figure 4 and simulated a multi-person communication scene using 3 tracks, 5 tracks, 7 tracks, and 9 tracks of speech audio. All the audio samples were processed through a Web-based automatic mixing system and through Web Audio recording to be collected and analysed.

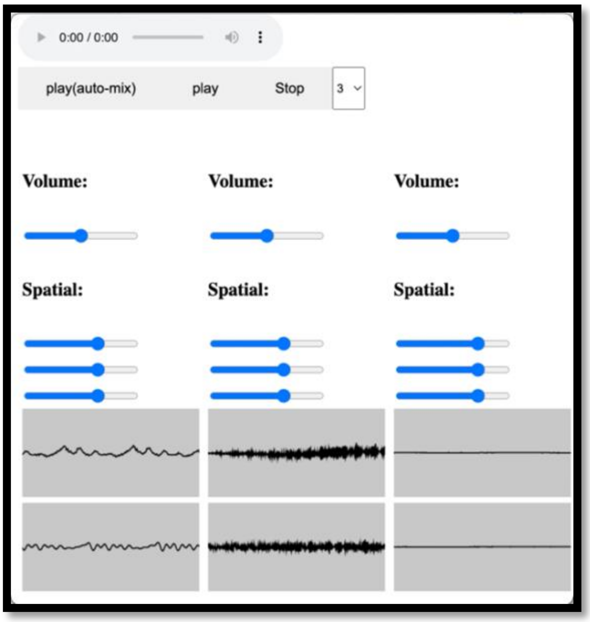


Figure 4: A screenshot of the proposed algorithm front-end on the Web.

4 Evaluation

We used an online listening test platform, Go Listen [27], to perform a multi-stimulus test to subjectively evaluate the effectiveness of the automatic unmasking system for multitrack content. The test methodology was similar to the Multiple Stimuli with Hidden reference (MUSHRA) [28].

We performed blind comparisons for mixes including our system, the original unmixed content (referred to as the hidden reference), a low range anchor, a middle range anchor, and a manual mix performed by an experienced audio engineer. All sound sources were first loudness normalized to prevent loudness differences from biasing the results.

The test designed took approximately 10-15 minutes to complete. To complete the test, participants were asked to use a computer, as well as headphones. There was in total of 17 participants in the test, demographic information of which can be found in Table 3.

During the test, participants were asked to compare the stimuli presented to them and give a score for the clarity of each audio, as well to recognize the number of different speech signals present.

Gender	Female	4
	Male	13
Audio experiences	Some	2
	Yes	15
Hearing impairment	No	17
	Yes	0

Table. 3: Demographic characteristics of participants.

5 Results

Test1

In the first test, participants were asked to rate each mix and check whether they can hear different number of speech signals present in the audio. As seen in Figure 5, the manual mix received a consistently high rating. The rating of the Auto mix (the mix derived from our system) was consistently close to the manual mix in participant's ratings. In the scenario of 3 and 7 people, the rating of our system's mix is similar with the that of the manual mix. In the scenario of 5 people, our system's mix rating exceeded that of the manual mix. Overall, the mix generated by the automatic mixing system performed better than the unmixed version in different scenarios.

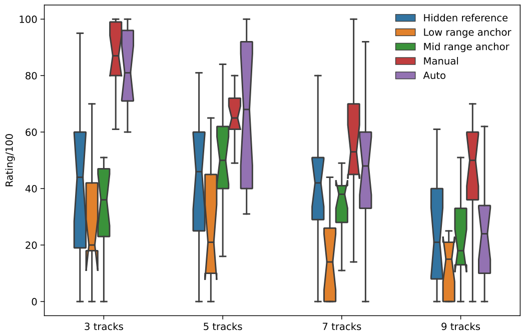


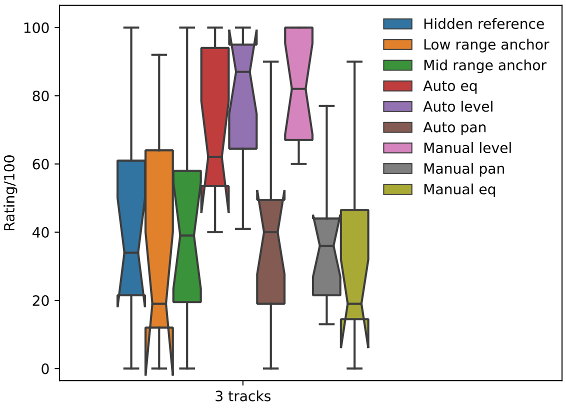
Figure 5: The multi stimulus test results comparing overall mixes.

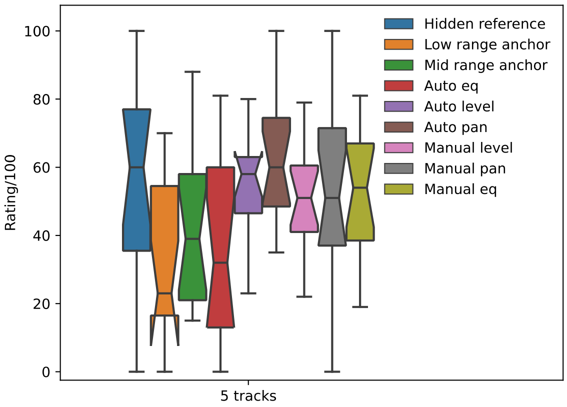
Test2

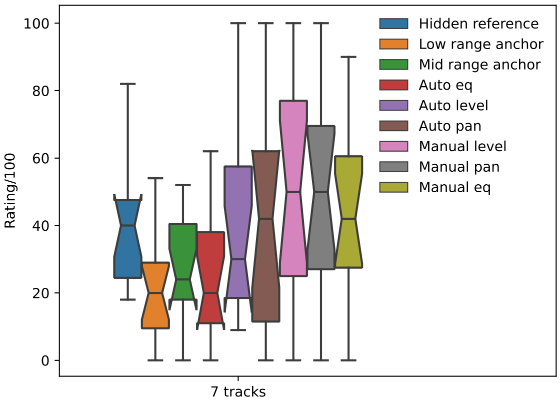
Our system was composed of three independent automatic mixing tools (EQ, spatialization, and level balance). Therefore, Test 2 asked participants to compare mixes using just one of those tools with equivalent manual mixes, see Figure 6. In the scenarios with 3, 7 or 9 tracks, the manual level balancing received the highest rating. For the 5 tracks scenario, the Auto Pan had the highest rating. From Figure 6, level adjustment is most effective for overall quality. The auto level and manual level received a good rating in each scenario.

The Auto EQ had a lower rating compared with other audio effects for 5, 7 and 9 tracks. Therefore, the Auto EQ may affect the result of the automatic mixing system's output quality and needs to be further evaluated and improved in future work.

Auto panning received a fairly good rating in the 5,7 and 9 people scenarios. However, both auto and manual panning received a low rating for 3 speakers, suggesting that panning may not be preferred when there are few sources.







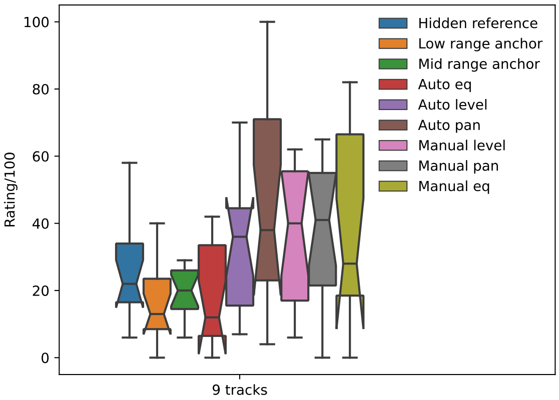


Figure 6: Subjective test results for different combinations of mixing tools, for 3, 5, 7 and 9 tracks.

Participants were again asked to rate the quality of the mix.

As expected, Figures 5 and 6 showed that the quality of the mix generally decreases with track count, most likely because intelligibility becomes increasingly difficult as the number of simultaneous speakers increases. This was true for both automatic and manual mixes.

Figure 6 also suggests that the level balance is the most important mixing tool for this task. This is due both to the fact that level imbalance directly causes masking of the lower amplitude track, and that the more frequency-specific masking issues are much harder to address with equalisation. However, this contrasts slightly with [7], which showed that panning was the most effective tool to address masking issues in their experiment.

Some participants reported that too many low frequencies can cause inaudibility, and some high frequency sibilance will affect the understanding of words. Additionally results showed that the clarity improves as the distance between speakers is increasing. In the 9 people scenario, participants could either concentrate and understand one voice, or all the voices together as "noise”.

6 Conclusions

In this work, we presented an online automatic mixing system which combined three audio effects to achieve unmasking of multiple-voice audio content. We compared our system’s work with original audio and manual mixing audio versions. Moreover, we explored the effectiveness of three different effects for mixing in speech scenarios. The level balancing appeared to be the most important factor for mixing multi-speaker scenarios.

Listening test ratings showed that overall, the automatic mixing system can compete with the manual mix and that level balancing is the most important factor for unmasking.

Future work should aim to find a more optimal solution for unmasking, and in particular, improve the automatic equalization tool. The system can be compared directly against existing microphone mixers, and scenarios where listeners prefer the ability to hear simultaneous speakers should be identified. It should also be deployed in real world scenarios, e.g., teleconferences, VR chat or multiplayer games.

Acknowledgements

This work was supported by the China Scholarship Council and Queen Mary University of London.

References

[1] Z. Wang et al., “NN3A: Neural Network Supported Acoustic Echo Cancellation, Noise Suppression and Automatic Gain Control for Real-Time Communications.” IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 661-665. (2021).

[2] B. C. Moore (ed.), An Introduction to the Psychology of Hearing. New York: Brill (2012).

[3] D. Howard and J. Angus, *Acoustics and Psychoacoustics*. Abingdon: Taylor &Francis (2013).

[4] B. Scharf, Critical bands. In Foundation of Modern Auditory Theory, vol. 1, pp. 159–202. (1970)

[5] B. C. Moore, et al. “A Model for the Prediction of Thresholds, Loudness, and Partial Loudness.” *Journal of the Audio Engineering Society*, vol. 45, no. 4, 1 Apr. 1997, pp. 224–240. (1977)

[6] S. Gelfand, Hearing: An Introduction to Psychological and Physiological Acoustics. New York: Routledge (2018)

[7] J. Wakefield and C. Dewey, “An Investigation into the Efficacy of Methods Commonly Employed by Mix Engineers to Reduce

Frequency Masking in the Mixing of Multitrack Musical Recordings," *138th AES Convention* (2015).

[8] B. De Man, R. Stables, and J. D. Reiss, *Intelligent Music Production*. New York, NY: Routledge, (2020).

[9] B. D. Man, J. Reiss, and R. Stables, "Ten Years of Automatic Mixing," 3rd Workshop on Intelligent Music Production, Salford, UK, (2017).

[10] E. Gonzalez and J. Reiss, "Improved Control for Selective Minimization of Masking Using Inter-Channel Dependancy Effects," *11th Intl.Conf. on Digital Audio Effects* (2008)

[11] A. Tom, J. D. Reiss, and P. Depalle, "An Automatic Mixing System for Multitrack Spatialization for Stereo Based on Unmasking and Best Panning Practices," *146th AES Convention*, (2019).

[12] Pedro D. Pestana and Joshua D. Reiss. A cross-adaptive dynamic spectral panning technique. In Proceedings of the 17th International Conference on Digital Audio Effects (DAFx-14), Erlangen, Germany, 2014.

[13] S. Hafezi and J. D. Reiss, "Autonomous Multitrack Equalization Based on Masking Reduction," *Journal of the Audio Engineering Society*, vol. 63, no. 5, pp. 312– 323, www.aes.org/e-lib/browse.cfm?elib=17637.(2015).

[14] D. Matz, E. Cano, and J. Abeßer, New Sonorities for Early Jazz Recordings Using Sound Source Separation and Automatic Mixing Tools. In ISMIR (pp. 749-755), (2015)

[15] P. Shekar and J. O. Smith, III, Modeling the harmonic exciter. In Proceedings of the 135th AES Convention, New York, USA, 2013.

[16] D. Dugan, "Automatic Microphone Mixing." *Journal of the Audio Engineering Society*, vol. 23, no. 6, pp. 442–449, (1975).

[17] Q. Wang et al., "VoiceFilter: Targeted Voice Separation by Speaker-Conditioned Spectrogram Masking." *ArXiv.org*, arxiv.org/abs/1810.04826 (2018).

[18] T. Zhou, Y. Zeng and R. Wang, Single-channel speech enhancement based on psychoacoustic masking. Journal of the Audio Engineering Society, 65(4), 272-284, (2017).

[19] R. Gu et al., "Multi-Modal Multi-Channel Target Speech Separation." *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 3, pp. 530–541 (2020).

[20] J. Wu et al., "Time Domain Audio Visual Speech Separation." IEEE automatic speech recognition and understanding workshop (ASRU), pp. 667-673 (2019).

[21] M. Rothbucher et al., "3D Audio Conference System with Backward Compatible Conference Server Using HRTF Synthesis." *J. Multim. Process. Technol* (2011).

[22] EBU R 128, "Loudness normalisation and permit- ted maximum level of audio signals," Recommendation, European Broadcasting Union. (2022).

[23] J. Reiss, *Working with the Web Audio API.* CRC Press, (2022).

[24] H. Zen et al., "LibriTTS: A Corpus Derived from LibriSpeech for Text-to-Speech," *arXiv.org*, https://arxiv.org/abs/1904.02882.(2019).

[25] E. Perez_Gonzalez, and J. Reiss, "An automatic gain normalisation technique with applications to audio mixing," 124th Audio Engineering Society Convention, (2008).

[26] A. de Cheveigné and H. Kawahara. "YIN, a Fundamental Frequency Estimator for Speech and Music." *The Journal of the Acoustical*

Society of America, vol. 111, no. 4, pp. 1917–1930, 10.1121/1.1458024. (2002).

[27] S. G. Kobourov, “Spring Embedders and Force Directed Graph Drawing Algorithms.” *ArXiv:1201.3011 [Cs]*, 14 Jan. arxiv.org/abs/1201.3011(2012).

[28] 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).

[29] International Telecommunication Union, “Multiple Stimuli with Hidden Reference and Anchor”, ITU-R BS.1534-1,（2003).