Asia-Pacific Journal for Arts Education

Special Issue

Unfold the Future of Music Education through Technology

Co-editors:
Prof. Bo Wah LEUNG
Dr. Cheung On TAM
Dr. Chi Hin LEUNG
Dr. Richard Guy WHITBREAD
The Education University of Hong Kong

http://www.ied.edu.hk/ccaproject/apjae/apjae.htm ISSN: 1683-6995

Volume 22 Number 3 December 2023

Deep Neural Networks with Music Dereverberation for Technical Ear Training in Music Production Education

Manni Chen

City University of Hong Kong

manni.chen@my.cityu.edu.hk

Prof. PerMagnus Lindborg

City University of Hong Kong

pm.lindborg@cityu.edu.hk

Shuo Meng

City University of Hong Kong

shuomeng2-c@my.cityu.edu.hk

Abstract

Reverberation is an audio effect used in music production affecting the audio spectrum, as well as the timbre, of music. Technical ear training concerning the ways that reverberation works on music samples is important, since in order to achieve the desired texture sound engineers need to be able to perceive subtle audio changes. However, “dry” recordings, i.e. those that lack reverberation, are not universally available for music production students for the purposes of practice. In this paper, we propose a deep learning-based music dereverberation method to generate de-reverbed music samples for technical ear training in music production education. The experiment results show that based on various objective evaluation metrics, the proposed method can effectively realize dereverberation compared to other neural network-based methods.

Key words

(de)reverberation, audio spectrum, timbre, texture, technical ear training, music production

Introduction

Music production involves several workflow stages in the form of recording, mixing and mastering (Reiss et al., 2019). Recording typically refers to the technique of using microphones to collect the sounds of instruments, vocals, etc. to store into a media format. Mixing concerns the balance of individual musical tracks in terms of a series of audio features, including timbre and special location, while mastering deals with the final polish to the mix before distribution. During the music production process, audio effects, defined as "the controlled transformation of a sound typically based on some control parameters" (Wilmering et al., 2020, p. 791), are crucial tools for forming music texture. According to Wilmering et al. (2013), audio effects can be classified by their perceptual attributes as loudness, duration and rhythm, pitch and harmony, space, and timbre or quality. The "loudness" group can also contain audio effects (compressors, for example, possesses parameters including thresholds, attack time, release time, and make-up), while the "space" group includes reverberation, where pre-delay, reverb time, and wet constitute some of the more familiar parameters.

Reverberation is an audio effect that is widely used in music production applications. It affects the timbre of music and influences the perception of sound space (Zölzer, 2011). In this case, applying reverberation appropriately is essential in terms of creating the appropriate timbral and spatial attributes. However, adding reverberation is about tuning a series of

parameters in order to control the audio transformation. Utilizing reverberation as a means of reaching the ideal perceptual target requires critical listening during the music production process, which means that sound engineers are required to make decisions on the setting of reverberation parameters based on their perceptual judgements. These perceptual judgements cover being aware of the indistinct musical discrepancies that exist when the parameters of audio effects are modified. In this respect, training in recognizing these almost unnoticeable changes is crucial for sound engineers. Consequently, the ability to understand how reverberation works on the perception of audio relies on a high level of technical ear training.

Technical ear training and reverberation

In traditional music education, ear training is a necessary skill for music composition and performance. In the field of music production, technical ear training is vital for work with audio signals through an ability to discern audio features. Corey & Benson proposed that technical ear training is a type of "perceptual learning focused on the timbral, dynamics and spatial attributes of sound" (2016, p. 5), and as such is an integral element of music production education.

Music production refers to the application of a series of technical tools to produce music. In addition to possessing a deep understanding of the theoretical background of sound production, sound engineers also need to be perceptually aware of music. In order to combine

the scientific and artistic contexts, critical listening skills, based on the subjective decisions of sound engineers in response to what they hear, can serve as a bridge linking technical knowledge and musical aesthetics. Specifically, the operations involved in music production, including balancing the mix and adding audio effects, rely heavily on the ability to apply listening skills to achieve the targeted texture by tuning a series of parameters on audio software and hardware devices. A person without the necessary level of technical ear training is not capable of differentiating the subtle details of sound. For example, hearing the differences between quicker and slower attack times in a compressor as a precursor to making a decision about the most suitable parameter is acritical listening skill frequently called upon in music production. The ability to make judgements based on perceptual differences and translate these into technical modifications as they relate to audio signals, thereby attaining the required music production goals, is thus a requisite for every sound engineer. In this sense, the practice of acquiring critical listening skills through technical ear training is an inevitable and vital attribute.

1.1 Digital reverberation

With the introduction of digital technology has come the digitisation of audio effects for contemporary music production. Digital Audio Effects (DAFx) are transformations that blend audio signals, the modifications being conducted through a series of control parameters (Chourdakis & Reiss, 2017). Reverberation refers to the reflections of the delayed and

attenuated copies of a direct sound (Zölzer, 2011); digital reverberation, as one of the DAFx, is a room acoustics modelling technique in which the technology is used to mimic those impulse responses. The responses are themselves divided into the direct sound, early reflections, and late reflections. The direct sound is the original sound triggered in the room, the reverberation occurring when the duplicates of the direct sound form as reflections. Defined by the directivity of the original sound and the physical properties of the surfaces, early reflections occur immediately after the sound event, while late reflections (perceived as decaying sound) are actually closely related to echoes. Accordingly, by mimicking the impulse responses in the room, digital reverberation, as adopted in music production, is able to bring the modifications of spaciousness, depth, and distances to the music, in the process creating different "environments".

Technical ear training can be helpful for music production students as a means of cultivating the ability to identify the changes brought about by reverberation. The learning materials for this ear training demand an ample number of music samples with reverberation effects (wet versions) and their corresponding dry recordings. However, it is unusual for the music production industry to release wet and dry version recordings at the same time (Martínez-Ramírez, 2022). Moreover, the recordings that *are* available are produced with added reverberation, which has led to a lack of (public) dry recordings. The result from the students' point of view is that merely listening to the reverbed version of a piece of music

will not enable them to distinguish between, and pass judgement on, how reverberation works and what kind of reverberation suits a particular style of music. For instance, the amount and type of reverberation added to Baroque and Romantic pieces of music are likely to be very different, partly because of the sonic environments and reflecting surfaces in architectural buildings from these two very different historical eras. If the only accessible ear training materials are produced music, students will have no idea how to properly add reverberation after recording Baroque music, since they are unable to perceive the distance between un-reverbed Baroque recordings and the corresponding wet versions. Conversely, for ethnic music and instruments, dry and clean recordings for music production are more difficult to obtain compared to wet and coloured versions, leading to a lack of familiarity with the un-reverbed textures of individual instruments. The combination of the types of requirements required for reverbed and un-reverbed audio samples and the challenges of finding ways to remove reverberation effects in music in order to expand the number of learning materials for technical ear training consequently provided the impetus for the deep learning methods introduced within the design of neural networks outlined in this research project.

Existing deep learning methods

In recent years, neural networks have started to be introduced into audio applications. Mimilakis et al. (2016) utilized deep neural networks and the short-time Fourier transform (STFT) for the purposes of predicting a coefficient of dynamic range compression in music

mastering applications, while the dilated residual network (DRN) (Grachten, 2018) and the convolutional neural network (CNN) (Ramírez & Reiss, 2018) have been applied to equalization in audio production. Oord et al. (2016) and Jang et al. (2021) proposed ‘WaveNet’ and ‘UnivNet’, respectively, for generating raw audio from text. However, the generative adversarial network (GAN) (Goodfellow et.al., 2020) has found the most favor among researchers in the field of audio application. Su et al. (2020) introduced ‘HiFi-GAN’, which combines ‘WaveNet’ and GAN for speech application, while Yamamoto et al. (2020) designed ‘Parallel WaveGAN’ for text-to-speech via the adoption of ‘WaveNet’ and a multi-resolution spectrogram. AlBadawy et al. (2022) proposed a neural vocoder which converts the spectral representations of audio signals into the generation of high-fidelity waveforms in real time. In an exception to the speech genre, GAN has also been utilized to inverse the reverberation caused by the car environment (Pepe et al., 2020), while Kumar et al. (2019) introduced ‘MelGAN’ for generating waveforms, including a synthesis of speech and music.

However, music production education in terms of technical ear training has been ignored by existing neural networks. The focus of this paper is on music dereverberation for the purposes of expanding the amount of technical ear training materials, based on the pre-existing condition that the construction of neural networks for music dereverberation can be applied to the same deep learning methods used in speech. Since the deep learning training

process is closely related to the spectrogram of audio materials, the differences between the frequency content of speech (used here to refer to the “standard” means of human vocal communication based around a narrow spectrum range and less complexity in terms of resonance compared to an individual’s singing voice) and music deserve to be highlighted. However, the presence of important differences between speech and music brings its own set of difficulties in terms of designing deep neural networks for music dereverberation. Additionally, some of the difficulties associated with dereverberation also lie within unknown acoustic environments and moving sources, both of which involve time-varying reverberation effects (Nercessian & Lukin, 2019).

As a result, we propose a GAN-based model with a ‘UNet’ (Ronneberger et.al., 2015) to achieve high-quality music dereverberation. First, the reverbed music samples are converted to full-band Mel spectrograms (Stevens et al., 1937) as input to gain frequency information. Then, a multi-scale ‘UNet’-based architecture is adopted to generate non-reverb Mel spectrograms. Finally, a GAN-based vocoder (containing a generator with a series of transposed convolutions that produce input-sized waveforms and a multi-period waveform discriminator to grasp both spectral and temporal domain features) is designed to synthesize fine-grained waveforms as the output.

Proposed method

Due to high temporal resolution and the dependencies of audio data, modeling raw audio is

particularly challenging. Most current methods therefore adopt an indirect approach through the modeling of a lower-resolution representation (Kumar et al., 2019) in order to reconstruct high-resolution audio. Inspired by 'UnivNet' (Jang et al., 2021) and 'U-Net' (Ronneberger et al., 2015), the overall architecture presented here contains two parts: a 'UNet'-based de-reverberator and a GAN-based neural vocoder, which contains a generator and a discriminator. The model 'learns' from paired examples using mono reverbed and dry audio samples. The block diagram of the proposed model is displayed in Figure 1.

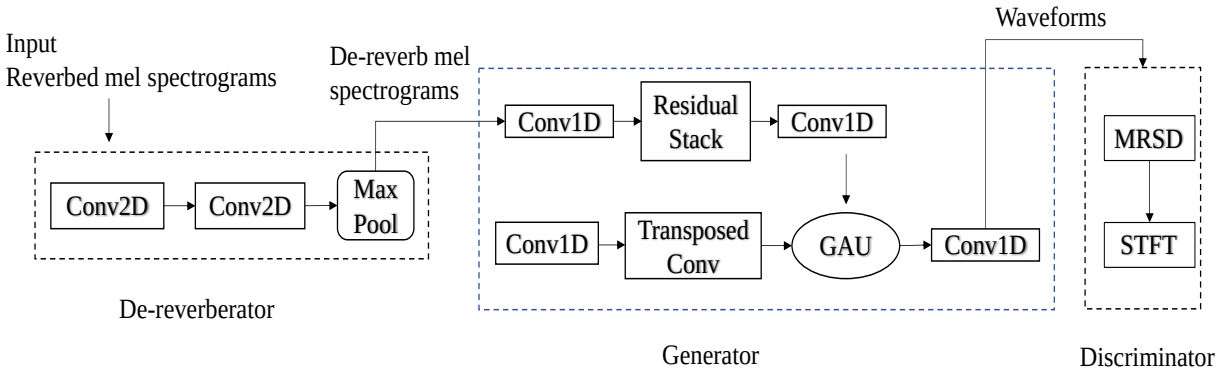

Figure 1 *Block diagram of the proposed model*

1.1 Architecture

'UNet' is widely used for image-to-image segmentation tasks (Ronneberger et al., 2015), dereverberation tasks (Ernst et al., 2018; Wang & Wang, 2020), and audio source separation tasks (Stoller et al., 2018). In this paper, we modify the 'UNet' structure for music dereverberation tasks; specifically, the architecture is a symmetric encoder-decoder network with skip connections which run on each encoder block and its mirror image between decoder

blocks. As for the input data, we transfer the waveform to a Mel spectrogram rather than an STFT to feed into the neural networks, since an STFT contains more data points and a Mel spectrogram corresponds to human perceptions of sound, both of which facilitate the generation of de-reverbed music. Compared to using raw audio waveforms, a Mel spectrogram is easier to model and retains enough information to convert back to audio (Kumar et al., 2019). All waveforms were resampled to 22.5kHz using an STFT with 1024 FFT components, 1024 sample Hann window lengths, and 0-12kHz Mel spectrograms with 80 bands. Input signals were normalized using the mean and variance of the entire training set. The final output of the de-reverberator is the same size as the input in order to predict the de-reverberated Mel spectrogram.

Directly applying a ‘UNet’-similar image-to-image architecture is hard to keep aligned with human perceptions. In this case, GAN-based neural vocoders are successfully applied to audio reconstruction tasks. Among them, ‘UnivNet’, one of the state-of-the-art neural vocoders, absorbed the advantages of various vocoders and used full band Mel spectrograms to obtain good speech reconstruction results. Consequently, this research adopted ‘UnivNet’ as the neural vocoder to conduct the music dereverberation tasks. The structure of the generator is the same as ‘MelGAN’, which used a background noise as the input and predicted the dry Mel spectrogram generated by the de-reverberator as the condition. Additionally, the location-variable convolution (LVC) and a gated activation unit (GAU)

were added and a multi-resolution spectrogram discriminator (MRSD) (Jang et al., 2021), which employs a multi-spectrogram and a multi-period waveform discriminator to differentiate dry recordings and generate audio samples, adopted. The converted multiple spectrograms contain various temporal and spectral resolutions, while the multi-period waveform enhances adversarial modeling in the temporal domain.

1.2 Training losses

For the de-reverberator, the mean square error (MSE) loss was used, while for the neural vocoder, multi-resolution STFT loss (Yamamoto et al., 2020) as an auxiliary loss L_{aux}, and the commonly used least-squares GAN as the objective functions, were applied. The auxiliary loss L_{aux} combines the log STFT magnitude loss L_{mag} and spectral convergence loss L_{sc}. To improve the results, we added the MSE loss to the neural vocoder as another regularization term in order to compare the differences between the de-reverberated audio samples and the dry audio samples. All loss functions were defined as follows:

$$L_{MSE} = \sum_t \|x_t - \hat{x}_t\|^2$$

$$L_{sc}(s,\hat{s}) = \frac{\|s-\hat{s}\|_F}{\|s\|_F},\ L_{mag}(s,\hat{s}) = \frac{1}{S}\|\log s - \log \hat{s}\|_1$$

$$L_{aux}(x,\hat{x}) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_{x,\hat{x}}\left[L_{SC}(s_m,\hat{s}_m) + L_{mag}(s_m,\hat{s}_m)\right]$$

$$L_{GAN}(G,D) = \sum_t \left(\log D\,(Z_t,X_t) + \log\left(1 - D\big(Z_t,G(Z_t)\big)\right)\right)$$

$$L(G,D) = L_{GAN}(G,D) + \lambda L_{MSE}(G)$$

Experiments

1.1 Datasets

The dry audio samples were downloaded from the Chinese Bamboo Flute (CBF) dataset (Wang et al., 2019) and 'The "Mixing Secrets" Free Multitrack Download Library'. The CBF dataset is a collection of 20 dry recordings of four Chinese bamboo flute pieces. As for the 'Mixing Secrets' dataset, in order to make sure that the audio samples contained more sustainable signals instead of silence as a means of facilitating the training process, multitrack excerpts, as opposed to the full multitrack, were used. Within the excerpt multitrack, we chose pop music and the dry recordings of songs from the multitrack, listening to every sample carefully in order to filter out the tracks that lacked continuous content, e.g., silence over three seconds, along with low-quality audio samples that included obvious artificial audio effects, such as pitch modification and lower clarity or breathing noise. The music samples were then cropped into blocks of three seconds each and the selected data separated into a training set consisting of 2400 audio samples and a testing set containing 600 randomly assigned audio pieces. Reverberation was added to the dry recordings via 'Pedalboard', a Python library containing the functionality of adding a series of audio effects. The built-in reverberation in 'Pedalboard', with parameters of a room size setting at 0.25, was adopted in order to make paired reverberated audio samples.

1.2 Evaluation metrics

In order to evaluate the performance of different models for dereverberation tasks, three objective metrics (Frequency-Weighted Segmental Signal-to-Noise Ratio (FWSegSNR), Log-Likelihood Ratio (LLR), and Signal to Distortion Ratio (SDR)) were employed. FWSegSNR is widely used for evaluating dereverberation performance (Kinoshita et al., 2016), defined as the average SNR ratio within a certain frequency band, the quality increasing with the value. LLR (Hu & Loizou, 2007) is also a conventional metric based on the LPC spectrum; suitable for the evaluation of reverberation effects, whereby those with smaller values are regarded as being of a better quality. SDR (Févotte et al., 2005) is a comprehensive index used to measure the sound quality of audio sources; the larger the value, the better the performance.

Results

The models ‘MelGAN’, ‘Unet+MelGAN’ and ‘Unet+UnivNet-c32’ were prepared for comparison. Since the original ‘MelGAN’ and ‘UnivNet’ act as neural vocoders, aimed at converting the Mel spectrograms into waveforms rather than dereverberation tasks, we added a cascade ‘UNet’ before them in order to realize the dereverberation. The cascade ‘UNet’ module took the Mel spectrogram of the reverbed audio as input to predict the dry Mel spectrogram. At the same time, we used ‘MelGAN’ to undertake the dereverberation tasks directly; this involved taking the Mel spectrogram of the reverbed sample as the input and

predicting the dry audio waveforms directly. All models were based on the official implementations and the relevant hyper-parameters following the reference configurations of each one. To ensure stable sound quality, the consumption times were calculated on the same platform and all deep learning-based models were trained on NVIDIA RTX3090 GPUs until the loss tended to stabilize. The proposed model was trained with a batch size of 32 and an AdamW optimizer (Loshchilov & Hutter, 2018), with $\beta1 = 0.5$, $\beta2 = 0.9$ and a 1e-4 learning rate.

The results of the objective comparisons are summarized in Table 1. They show that using 'MelGAN' to generate de-reverbered audio samples directly failed to obtain ideal results, although it was the fastest model. By comparison, the neural network-based methods using 'UNet' cascaded resulted in a similar performance, illustrating that an indirect model is a requisite. The proposed method reaches the highest performance among the compared models in all objective metrics, except for speed, which can be tolerated in the real scenarios. Overall, the results demonstrate that the proposed method can realize high-quality music dereverberation while almost staying in real time.

Table 1 *Comparisons between different methods*

Model Name	**FWSegSNR**	**LLR**	**SDR**	**Speed**
'MelGAN'	6.57	1.03	7.80	**177.51**

‘Unet + MelGAN’	9.09	0.51	9.17	142.86
‘Unet + UnivNet’	10.58	0.44	9.14	118.11
Proposed Method	**12.31**	**0.29**	**9.67**	98.04

Note. The speed is calculated based on consumption time relative to real time.

Conclusions

In music production education, technical ear training is ineluctable when comparing un-reverbed and reverbed music. Moreover, the lack of un-reverbed samples has resulted in the obstacle of music production students failing to understand comprehensively how reverberation works. In order to tackle this problem, while simultaneously expanding un-reverbed music learning materials, this paper has proposed a GAN-based model to generate real time and high quality dereverberation music. The results of the different evaluation metrics indicate that our method is more capable of realizing music dereverberation compared to other neural network-based methods. Additionally, this implementation can be extended to other audio effects, such as delay, in the process exploring audio transformations as they pertain to technical ear training. Future explorations will include more universal, reality-sited dereverberation models based on unparalleled data, together with the extension of the applications to other audio effects in order to facilitate

technical ear training within music production education.

References

AlBadawy, E. A., Gibiansky, A., He, Q., Wu, J., Chang, M. C., & Lyu, S. (2022, May). Vocbench: A Neural Vocoder Benchmark for Speech Synthesis. In ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 881-885). IEEE.

Corey, J., & Benson, D. H. (2016). *Audio production and critical listening: Technical ear training*. Routledge.

Ernst, O., Chazan, S. E., Gannot, S., & Goldberger, J. (2018, September). Speech dereverberation using fully convolutional networks. In 2018 26th European Signal Processing Conference (EUSIPCO) (pp. 390-394). IEEE.

Foote, C., Gribonval, R., & Vincent, E. (2005). BSS_EVAL toolbox user guide—Revision 2.0.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2020). Generative adversarial networks. *Communications of the ACM*, *63*(11), 139-144.

Grachten, M., Deruty, E., & Tanguy, A. (2018). Auto-adaptive resonance equalization using dilated residual networks. arXiv preprint arXiv:1807.08636.

Hu, Y., & Loizou, P. C. (2007). Evaluation of objective quality measures for speech enhancement. *IEEE Transactions on audio, speech, and language processing*, *16*(1), 229-238.

Jang, W., Lim, D., Yoon, J., Kim, B., & Kim, J. (2021). UnivNet: A neural vocoder with multi-resolution spectrogram discriminators for high-fidelity waveform generation. arXiv preprint arXiv:2106.07889.

Kinoshita, K., Delcroix, M., Gannot, S. P., Habets, E. A., Haeb-Umbach, R., Kellermann, W., Leutnant, V., Maas, R., Nakatani, T., Raj, B., Sehr, A., & Yoshioka, T. (2016). A summary of the REVERB challenge: state-of-the-art and remaining challenges in reverberant speech processing research. *EURASIP Journal on Advances in Signal Processing*, *2016*(1), 1-19.

Kumar, K., Kumar, R., de Boissiere, T., Gestin, L., Teoh, W. Z., Sotelo, J., & Courville, A. C. (2019). Melgan: Generative adversarial networks for conditional waveform synthesis. Advances in neural information processing systems, 32.

Loshchilov, I., & Hutter, F. (2018). Fixing weight decay regularization in Adam.

Martínez-Ramírez, M. A., Liao, W.-H., Fabbro, G., Uhlich, S., Nagashima, C., & Mitsufuji, Y. (2022). Automatic music mixing with deep learning and out-of-domain data. In the 23rd International Society for Music Information Retrieval Conference (ISMIR). https://doi.org/10.48550/arxiv.2208.11428

Mimilakis, S. I., Drossos, K., Virtanen, T., & Schuller, G. (2016, May). Deep neural networks for dynamic range compression in mastering applications. In Audio Engineering Society Convention 140. Audio Engineering Society.

Nercessian, S., & Lukin, A. (2019, September). Speech dereverberation using recurrent neural networks. In Proceedings of the 23rd International Conference on Digital Audio Effects (DAFx-19), Birmingham, UK (pp. 2-6).

Oord, A. V. D., Dieleman, S., Zen, H., Simonyan, K., Vinyals, O., Graves, A., & Kavukcuoglu, K. (2016). Wavenet: A generative model for raw audio. arXiv preprint arXiv:1609.03499.

Pepe, G., Gabrielli, L., Squartini, S., & Cattani, L. (2020). Designing audio equalization filters by deep neural networks. *Applied Sciences*, *10*(7), 2483.

Ramírez, M. A. M., & Reiss, J. D. (2018, September). End-to-end equalization with convolutional neural networks. In 21st International Conference on Digital Audio Effects (DAFx-18).

Reiss, J., Stables, R., & De Man, B. (2019). *Intelligent Music Production: A Theoretical Overview*. Routledge.

Ronneberger, O., Fischer, P., & Brox, T. (2015, October). U-net: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention (pp. 234-241). Springer, Cham.

Stevens, S. S., Volkmann, J., & Newman, E. B. (1937). A scale for the measurement of the psychological magnitude pitch. *The Journal of the Acoustical Society of America*, *8*(3), 185-190.

Stoller, D., Ewert, S., & Dixon, S. (2018). Wave-u-net: A multi-scale neural network for end-to-end audio source separation. arXiv preprint arXiv:1806.03185.

Su, J., Jin, Z., & Finkelstein, A. (2020). HiFi-GAN: High-fidelity denoising and dereverberation based on speech deep features in adversarial networks. arXiv preprint arXiv:2006.05694.

Wang, C., Benetos, E., & Chew, E. (2019, November). CBF-PeriDB: A Dataset of Chinese Bamboo Flute Playing Techniques with Periodic Modulations. In 20th International Society for Music Information Retrieval Conference.

Wang, Z. Q., & Wang, D. (2020). Deep learning-based target cancellation for speech dereverberation. *IEEE/ACM transactions on audio, speech, and language processing*, *28*, 941-950.

Wilmering, T., Fazekas, G., & Sandler, M. B. (2013, October). Audio effect classification based on auditory perceptual attributes. In Audio Engineering Society Convention 135. Audio Engineering Society.

Wilmering, T., Moffat, D., Milo, A., & Sandler, M. B. (2020). A history of audio effects. *Applied Sciences*, *10*(3), 791.

Yamamoto, R., Song, E., & Kim, J. M. (2020, May). Parallel WaveGAN: A fast waveform generation model based on generative adversarial networks with multi-resolution spectrogram. In ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 6199-6203). IEEE.

Zhu, J.-Y., Park, T., Isola, P., & Efros, A. A. (2017). Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. 2017 IEEE International Conference on Computer Vision (ICCV), 2242–2251. IEEE. https://doi.org/10.1109/ICCV.2017.244

Zölzer, U. (2011). DAFX: Digital Audio Effects, Second Edition (2nd ed.). Chichester, West Sussex, U.K: Wiley. The ‘Mixing Secrets’ Free Multitrack Download Library. Retrieved from https://www.cambridge-mt.com/ms/mtk/