# ACOUSTIC MATCHING BY EMBEDDING IMPULSE RESPONSES

*Jiaqi Su* [†⋆] *Zeyu Jin* [⋆] *Adam Finkelstein* [†]

[†]Princeton University [⋆]Adobe Research

## ABSTRACT

The goal of acoustic matching is to transform an audio recording made in one acoustic environment to sound as if it had been recorded in a different environment, based on reference audio from the target environment. This paper introduces a deep learning solution for two parts of the acoustic matching problem. First, we characterize acoustic environments by mapping audio into a low-dimensional embedding invariant to speech content and speaker identity. Next, a waveform-to-waveform neural network conditioned on this embedding learns to transform an input waveform to match the acoustic qualities encoded in the target embedding. Listening tests on both simulated and real environments show that the proposed approach improves on state-of-the-art baseline methods.

***Index Terms*—** Acoustic Matching, Acoustic Impulse Response, Equalization Matching, Embedding, Reverberation

## 1. INTRODUCTION

Audio recordings play a central role in various media such as movies, voice-overs, audio books, podcasts, and vlogs. While many recordings are made in professional studios, a large portion of creative content relies on audio recorded in natural spaces, resulting in a wide variety of noise and reverberation as part of the recordings. Content created by amateur users adds another layer of variation in the quality of recorded audio, due to consumer grade devices and recording setups. As a result, reusing and merging such content can be difficult. Moreover, with the emergence of text-based audio editing [1], directly inserting new or synthesized speech content into an existing real-world recording can sound (literally) out of place. Another example scenario where audio matching is needed is in movie making, where dubbed speech recorded in the studio is later inserted into a scene to replace the original speech, a process often referred to as Automated Dialog Replacement (ADR).

Such problems can be addressed by acoustic matching, a process that transforms recordings made in one environment (the *source*) to match a new environment (the *target*), characterized by samples recorded in the target environment. The goal is to make the source sufficiently similar to the target so that when they are stitched together, the difference is unnoticeable. While previous work has addressed aspects of this problem via matching equalization (EQ) [2], our work addresses the general acoustic matching problem including reverberation, EQ distortion and noise for single-channel audio recordings.

In this paper, we propose a generic one-shot acoustic matching method based on deep learning. By *generic* we mean that the method is independent of speaker, content and source acoustic environment. By *one-shot* we mean no learning is required to adapt to a new environment; instead, an example is used to characterize the target. Our solution comes in two parts. First, we model an acoustic embedding that extracts the characteristics of a recording environment from speech recordings, and condenses it into a low-dimensional representation. Next, we train a waveform-to-waveform (end-to-end) neural network that transforms speech recordings in a source environment to a target, as indicated by the acoustic embedding of an example target recording. We build this model on the DAPS dataset [3] as well as a collection of room impulse responses and noise recordings [4, 5, 6]. Our listening tests show significant improvement over previous work, for both real and synthetic acoustic matching scenarios. Thus, our contributions are:

1. An embedding space for acoustic impulse responses independent of speaker and speech content.
2. A generic one-shot waveform-to-waveform acoustic matching network based on this embedding.
3. A simple and high-quality clean-to-environment matching solution based on nearest neighbor search in the embedding space.
4. A human listening study over a variety of alternative approaches and including both real and synthetic environments.

## 2. RELATED WORK

Resarchers addressing the ACE challenge [6] explore blind estimation of two acoustic environment parameters from recorded speech: direct-to-reverb ratio (DRR), which describes the energy ratio of direct arrival sound and reflected sound, and reverberation time (RT60), which describes the time it takes for a sound to decay 60dB. For single-channel estimation, statistical analysis of subband information [7, 8] and neural networks on spectral features [9, 10] have been employed. In general, predicting acoustic parameters is a difficult task. Given DRR and RT60, room impulse responses (RIRs) for reverberation can be modeled as white noise modulated by an exponentially decaying envelope [11], but this simple model does not capture subtleties such as early reflection patterns and coloration.

There are efforts as well for directly estimating acoustic responses from recorded speech. Since the problem is ill-posed, the majority of approaches rely on knowing emitting source statistics [12] or having multiple channels [13]. Non-negative matrix deconvolution and non-negative matrix factorization [14, 15] on spectrograms also estimates acoustic responses as a side product of de-noising and de-reverberation.

A parallel line of research focuses on generating artificial acoustic responses based on a few control parameters to produce realistic perceptual effects. The image method [16] is widely used for simulating reverberation from rectangular room geometries. In the emerging domain of augmented listening for VR and AR, the method of Li et al. [17] maps out room geometries from 360-degree videos and simulates reverberation for scene-aware audio by composing a synthesized early reflection with a measured late reverberation tail. However, there remains a performance gap between synthetic reverberation and real reverberation, as real environments can have more complicated spatial configurations and acoustic properties.

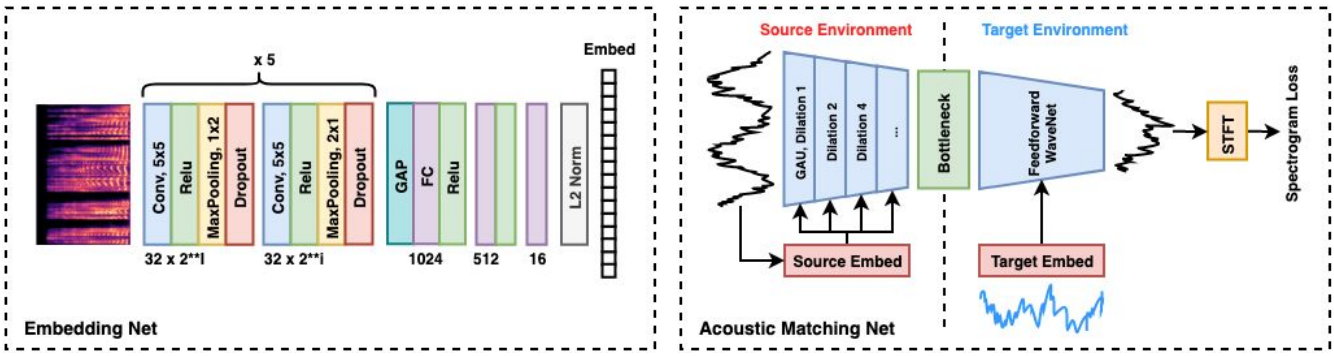


**Fig. 1**: Network Architecture. The embedding network takes in the log spectrogram of a recording and outputs an embedding vector of fixed dimension. The acoustic matching network is a stack of two feed-forward WaveNets globally conditioned on the outputs of the embedding network, and with convolutional bottleneck layers that project down the channel dimension in between. It takes in the input waveform of a source environment as well as a reference waveform (in blue) of a target environment, and maps to a waveform in the target environment.

Several real impulse response datasets are collected to account for the differences between real environments and simulated ones, such as the MIT Impulse Response Survey Dataset [4], the REVERB Challenge database [5] and the ACE Challenge database [6] – typically of hundreds of RIRs. Convolving such measured RIRs with clean speech produces reverberant speech. However, gathering new RIRs is expensive because of the specific recording setup required, and thus it is impractical to measure acoustic responses of recording environments from users' devices at application time. To expand existing RIR datasets, Bryan [18] proposes a mathematically viable augmentation to RIRs by re-scaling their DRR and RT60 properties, and Kearney et al. [19] investigates interpolation between RIRs via dynamic time warping.

The closest problem to acoustic matching is equalization matching for spectral balance. Mathur et al. [20] use CycleGAN to learn a mapping of sound between different microphones. Ramírez and Reiss [21] learn an end-to-end model to approximate the equalization target as a content-based transformation. Germain et al. [2] address mismatched coloration and background noise between different environments, and propose source-differentiated equalization matching that matches speech and noise separately. The equalization matching approaches, however, do not address cases where reverberation sounds different between environments. Thus, matching all of reverberation, equalization and noise remains an open problem.

## 3. APPROACH

We assume that reverberation and equalization are convolutional and that noise is additive, as summarized in the following equation:

$$\begin{aligned} y &= g * (h * s + n) \\ &= (g * h) * s + (g * n) \\ &= h' * s + n' \end{aligned}$$

where $s$ is clean speech, $h$ is the room impulse response for reverberation, $g$ is distortion introduced by recording devices and post-processing as a series of linear filters, $n$ is background noise, and $y$ is resulting recorded speech. Reverberation and equalization filters can be further combined into one acoustic impulse response $h'$, and the same goes for noise $n'$. Given speech recorded in a specific environment $y_s = h'_s * s_s + n'_s$, we aim to apply an end-to-end neural network that removes noise $n'_s$ and alters the acoustic impulse response to match the target environment $h'_t$ characterized by an embedding of the target environment $E(h'_t)$. Then we add target environment's noise $n'_t$ to obtain $y_t$.

### 3.1. Acoustic Environment Embedding

The idea of using environment characterization as auxiliary information has been shown to improve accuracy for speech recognition. Giri et al. [22] use estimated DRR and RT60 as conditions, while Pironkov et al. [23] estimate an i-vector representing speaker and acoustic environment. Kim et al. [24] learn an environmental noise embedding using neural network bottleneck. For acoustic matching, the accuracy of the embedding as indication of target environments is crucial, and thus we aim to model a space where distance in the embedding directly correlates to how different the two environments sound. This prompts us to learn an acoustic impulse response embedding space pre-trained from semi-hard triplet loss [25].

Our embedding network takes in the log spectrogram of a noisy reverberant recording of arbitrary length and outputs a fixed dimension encoding (Figure 1-left). It is composed of stacks of 2D convolutions, ReLU, max pooling and dropout with doubled channel dimensions, followed by a global average pooling across both temporal and frequency axes, three fully connected layers with ReLU activation, and L2 normalization at the end. Each triplet is constructed based on the class identities of the room impulse responses in our dataset, and noise of various types is randomly added to the samples. The training objective is to place recording samples with same acoustic impulse response closer and those of different acoustic impulse responses further, regardless of the noise present. The final embedding dimension is determined based on validation accuracy on environment classification task. We would like the embedding space to be as compact as possible while preserving accuracy, and found empirically that 16 dimensions is sufficient for characterizing different impulse responses.

### 3.2. Acoustic Matching Network

We employ a neural network to approximate the complicated mapping from recordings of source environment to recordings of target environment. Our architecture is based on feed-forward WaveNet, which is a waveform-to-waveform model with non-casual dilated convolutions and has been successfully applied to speech denoising [26, 27] and de-reverberation [28]. It avoids the issue of phase inversion by operating in the time domain. The dilated convolutions with exponentially increasing dilation rates enable a large receptive field that covers the length of typical impulse responses (thousands of samples).

To adapt to the acoustic matching problem, our network is designed as a stack of two feed-forward WaveNets with bottleneck layers in between (Figure 1-right). The first feed-forward WaveNet is

designed to strip away the source environment information and extract the speech content, while the second feed-forward WaveNet is to inject the target environment information by filtering. The bottleneck layers, as a series of 1D convolutions applied to the aggregated outputs of the skip connections of the WaveNet, project down the channel dimension of the tensors to force content extraction while preserving temporal resolution. The original WaveNet [29] incorporates global conditioning to guide the network to produce audio with the required characteristics across all time steps. For our task, we feed the embedding of the source environment and that of the target environment, inferred from the input recording and a reference recording of the target environment via our embedding network, as global conditioning respectively to both feed-forward WaveNets.

Since our acoustic matching network produces real-valued outputs, it learns a deterministic mapping function that does not model randomness well. Therefore, the training objective is to match only reverberation and EQ distortion. Noise is added separately after the neural network maps the input recording to the noise-free version of the target environment.

We find it is more effective during training to use the perceptually-motivated spectrogram loss proposed by Su et al. [28], rather than simple L1 or L2 loss, as sample-level differences are less meaningful in the presense of heavy reverberation or distortion. In practice, we use a equally weighted combination of two spectrogram losses with two sets of STFT for the sampling rate of $16\,\mathrm{kHz}$: one with large FFT window size of 2048 and hop size of 512, and one with small FFT window size of 512 and hop size of 128. The larger one gives more frequency resolution, while the smaller one gives more temporal resolution. We find a fixed pre-trained environment embedding space leads to over-fitting to known environments. Thus, we co-train the embedding network and the acoustic matching network, using data augmentation, so that the models have more flexibility to generalize. Meanwhile, the reference recording is randomly sampled so that it does not correlate with the content of the input.

### 3.3. Data Augmentation

To generalize to new speakers, new speech content, and new environments, we rely on several data augmentation techniques that introduce variation in the data on the fly during training. Speech is re-sampled at a random rate (90-110%), with randomly scaled amplitude (50-150%). Noise is chosen from a sample collection, passed through a random multi-band filter, and added with a random SNR ($10$-$30\,\mathrm{dB}$). For each room impulse response, we adjust its DRR by randomly scaling its direct signal response according to Bryan's proposed procedure [18], and adjust its RT60 by stretching or shrinking via re-sampling. We also apply random multi-band filters to the impulse responses as EQ distortion: 0-50 Hz, 50-300 Hz, 300-1500 Hz and 1500-8000 Hz, with random gain ($\pm 10\,\mathrm{dB}$).

Note that over-fitting persists despite this heavy data augmentation, showing that our dataset has limited coverage over the environment embedding space. This motivates future efforts to gather broader environment data as well as other augmentation methods.

### 3.4. Nearest Neighbor IR

As an alternative to the end-to-end acoustic matching approach, the embedding space we learnt provides a distance metric between environments. Thus we can look for the nearest neighbor acoustic impulse response as an approximation to the environment from a reference recording. In practice, we search among the embedding of our augmented dataset and choose the closest. This approach is applicable to scenarios where clean recordings are available. For example, when dubbing for movies, the retrieved impulse response can be convolved with a clean studio recording to match a target environment. Moreover, the acoustic matching network has a fixed receptive field, which limits the length of the reverberation tail it can model; a longer reverberation thus requires a deeper network which can be costly. In contrast, the nearest neighbor approach offers an efficient solution for reverberation of arbitrary length.

### 3.5. Matching Noise

The acoustic matching network implicitly de-noises the input recording, and thus noise matching the target environment should finally be added back in. We assume background noise is relatively stationary, and there is silence inside the target reference recording (in our case, several seconds at the beginning and end). Thus we model the target noise as filtered white noise with matching log spectrograms.

## 4. EVALUATION

We evaluate our method using a stack of two 10-layer feed-forward WaveNets for acoustic matching. The channel size is 128 across the entire network, except at the bottleneck which is composed of three convolutional layers of $3\times1$ filters at 16 channels. The STFT for the embedding network uses a window size of 1024 and a hop size of 512, at a sampling rate of $16\,\mathrm{kHz}$. We jointly train the acoustic matching network and the embedding network for a million iterations with a batch size of 10. The ADAM optimizer is used with learning rate 0.0001, reduced by a factor of ten after 500K iterations.

The speech corpus comes from the studio quality recordings of the DAPS Dataset's clean set [3], and is convolved with the 270 diverse impulse responses of the MIT Impulse Response Survey Dataset [4]. Noise is drawn from the REVERB Challenge database [5] and the ACE Challenge database [6]. This gives us parallel recordings for pairs of environments. In total, the data covers two genders $\times$ ten speakers per gender $\times$ ten minutes of script $\times$ 270 environments before augmentation. We hold out two minutes of script ("script5") from a speaker of each gender ("f10" and "m10"), as well as 70 environments ("h201" and above), for evaluation, while the remaining data is used in training. Our evaluation set also includes real-world recordings from the DAPS dataset – recordings of the same held-out speech under different room environments. They are produced by replaying the clean studio recordings in typical rooms and re-recording them with consumer devices, so that interactions of acoustic factors in real world are captured. We exclude the room "balcony", which contains strong noise, and use the remaining ten rooms in our evaluation.

There are no existing objective metrics for acoustic similarities that accurately model human perception. Therefore, our evaluations are based on listening tests via Amazon Mechanical Turk (AMT), a crowd-sourcing platform commonly used for such experiments [30]. We create test audio clips by stitching two consecutive utterances from two different environments and apply our methods to one of the two to acoustically match the other. If they match well, they should sound like a single, seamless recording; otherwise, a transition will be heard. In each human intelligence task (HIT), a subject is instructed to wear headphones and presented with 12 questions. In each question, the subject is asked to rate how seamless the audio clip sounds on a scale from 1=*very different* to 5=*seamless*. Four additional questions have obvious answers, and permit us to exclude ratings from subjects who do not understand the task.

We collected data from 3000 HITs, consisting of three sets of environment pairs to evaluate different aspects of the approaches, including clean-to-simulated noise-free environments (18 pairs),

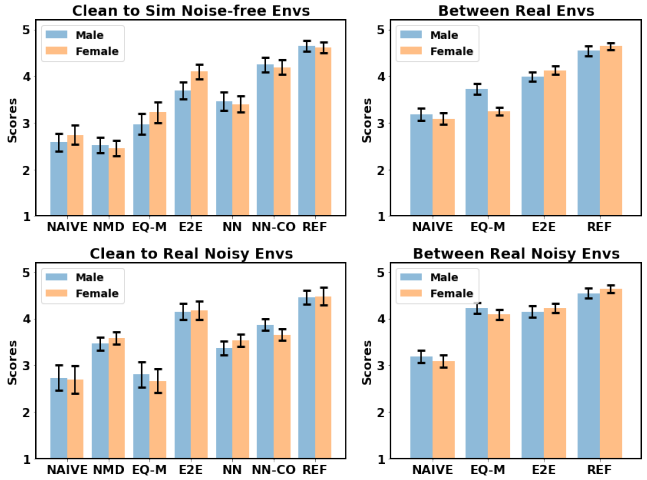


**Fig. 2**: Subjective ratings for four sets of environment pairs, across speaker gender and conditions. Error bars denote 95% confidence.

clean-to-real noisy environments (10 pairs), and real-to-real environments (28 pairs). For each transformation, we used 20 utterances per held-out speaker for creating test audio clips. Each test audio clip was evaluated by 16 subjects. Subjects evaluated three variants of our approach and two baseline methods from the literature. These are also calibrated relative to a naïve method (NAIVE), which does no acoustic matching, and a reference ground truth answer (REF):

1. E2E: Our end-to-end acoustic matching network.
2. NN: Our nearest neighbor IR retrieved from the pre-trained embedding space and applied to clean speech.
3. NN-CO: Our nearest neighbor IR retrieved from the co-trained embedding space and applied to clean speech.
4. EQ-M: Source-differentiated equalization matching [2].
5. NMD: IR estimated via non-negative matrix deconvolution [15] using our training data as exemplars, and applied to clean speech.

All these audio samples can be found at our project website.[1]

### 4.1. Noise-free Acoustic Impulse Response Evaluation

To evaluate the quality of the IR embedding solely, we conducted experiments of matching clean recordings to noise-free recordings simulated by convolving the randomly selected held-out impulse responses (from the MIT IR Survey Dataset) with the held-out clean speech, so that they contain only reverberation and equalization. As shown in Figure 2-top-left, NN-CO performs best (close to REF) which demonstrates the effectiveness of our learnt environment embedding space as a similarity metric. Co-training the embedding network with acoustic matching significantly improves over NN. Meanwhile, E2E gives the second best performance, but worse than NN-CO. We hypothesize two reasons: first, the E2E network introduces artifacts for the hard cases (high EQ distortion and strong reverb), likely due to the network's limited capability to handle corner cases. Second, the E2E network assumes a fixed receptive field that is smaller than the length of the tail of heavy reverb. This causes the network to generate a different sounding reverb when conditioned on long-tail reverb. In contrast, NN-CO does not exhibit artifacts or have a limited receptive field, and therefore produces speech that sounds almost the same as the ground truth.

[1] https://pixl.cs.princeton.edu/pubs/Su_2020_AMB/

### 4.2. Clean-to-environment Acoustic Matching Evaluation

The second study tests our method's ability to identify impulse response embedding correctly in the presence of noise. The goal is to match a clean recording to a target environment recording so that it sounds like the target environment with noise, reverberation and equalization all matched. This experiment is conducted on the real-world recordings of the DAPS dataset, where the clean set is used as input and all the other room categories are used as target. The lower left plot of Figure 2 depicts the result. With the presence of noise, the NN-based approach starts to degrade but is still significantly better than baseline EQ-M. This is likely due to noise interfering with IR embedding, causing a small variation in the space and thus a different IR being selected by nearest neighbor. NMD achieves a better performance than before, likely because it is designed for de-noising and handles noise well. Remarkably, the performance of E2E is increased. Our hypothesis is that E2E co-trained with IR embedding also learns to encode and re-generate noise; the noise in the target also helps mask other types of degradation, making minor artifacts produced by E2E less noticeable and hence increasing perceptual similarity to the target environment.

### 4.3. Full Acoustic Matching Evaluation

Finally, we conducted a study on the generic acoustic matching task: between real environment conditions in the DAPS dataset. For each voice, we randomly selected 28 pairs out of all possible source and target environments. Note we omit NN-based methods and NMD from this study, as we no longer have clean audio as input. Results appear in Figure 2-top-right. While E2E achieves consistent performance as before, EQ-M has significantly improved. A closer look at the results suggests that our method performs better than EQ-M in cases where two environments have bigger differences, especially in cases where a clean recording is matched to a noisy recording or vice versa. However, in a subset of cases where a noisy environment is matched to another noisy environment (Figure 2-lower-right), E2E maintains its quality while EQ-M becomes competitive. This is likely due to masking effect from the noise present in both the input and the target. It is also easier for EQ-M to filter existing reverb in the input than to add new reverb.

## 5. CONCLUSION

We present a deep learning method for acoustic matching that is able to handle novel speakers, speech content and environment acoustics, given relatively stationary background noise. A waveform-to-waveform acoustic matching neural network conditioned on (and co-trained with) an acoustic environment embedding learns to map recordings between different environments. The learnt low-dimensional embedding space identifies and characterizes acoustic impulse responses from recorded speech, where we can effectively approximate the target acoustic impulse response via nearest neighbor search in the augmented dataset. We show in subjective evaluations that our method improves significantly for hard cases where acoustic impulse responses are drastically different, especially between clean and noisy reverberant audio.

Potential areas for future work include modeling a noise embedding, and a generative method for realistic non-stationary noise based on the embedding. Multi-target learning could enforce desired constraints on the embedding spaces, so that different environment factors could be described independently.

## 6. REFERENCES


[1] Zeyu Jin, Gautham J Mysore, Stephen Diverdi, Jingwan Lu, and Adam Finkelstein, "Voco: Text-based insertion and replacement in audio narration," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 96, 2017.

[2] François G Germain, Gautham J Mysore, and Takako Fujioka, "Equalization matching of speech recordings in real-world environments," in *ICASSP 2016*, pp. 609–613.

[3] Gautham J Mysore, "Can we automatically transform speech recorded on common consumer devices in real-world environments into professional production quality speech? A dataset, insights, and challenges," *IEEE Signal Processing Letters*, vol. 22, no. 8, pp. 1006–1010, 2015.

[4] James Traer and Josh H McDermott, "Statistics of natural reverberation enable perceptual separation of sound and space," *Proceedings of the National Academy of Sciences*, vol. 113, no. 48, pp. E7856–E7865, 2016.

[5] Keisuke Kinoshita, Marc Delcroix, Takuya Yoshioka, Tomohiro Nakatani, Armin Sehr, Walter Kellermann, and Roland Maas, "The REVERB challenge: A common evaluation framework for dereverberation and recognition of reverberant speech," in *WASPAA 2013*, 2013, pp. 1–4.

[6] James Eaton, Nikolay D Gaubitch, Alastair H Moore, Patrick A Naylor, Eaton, et al., "Estimation of room acoustic parameters: The ACE challenge," *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, vol. 24, no. 10, pp. 1681–1693, 2016.

[7] Feifei Xiong, Stefan Goetze, Birger Kollmeier, and Bernd T Meyer, "Joint estimation of reverberation time and early-to-late reverberation ratio from single-channel speech signals," *IEEE/ACM TASLP*, vol. 27, no. 2, pp. 255–267, 2018.

[8] Thiago de M Prego, Amaro A de Lima, Rafael Zambrano-López, and Sergio L Netto, "Blind estimators for reverberation time and direct-to-reverberant energy ratio using subband speech decomposition," in *WASPAA 2015*, pp. 1–5.

[9] Pablo Peso Parada, Dushyant Sharma, Toon van Waterschoot, and Patrick A Naylor, "Evaluating the non-intrusive room acoustics algorithm with the ACE challenge," *arXiv preprint arXiv:1510.04616*, 2015.

[10] Hannes Gamper and Ivan J Tashev, "Blind reverberation time estimation using a convolutional neural network," in *2018 16th International Workshop on Acoustic Signal Enhancement (IWAENC)*, pp. 136–140.

[11] Katia Lebart, Jean-Marc Boucher, and Philip N Denbigh, "A new method based on spectral subtraction for speech dereverberation," *Acta Acustica united with Acustica*, vol. 87, no. 3, pp. 359–366, 2001.

[12] Dinei Florencio and Zhengyou Zhang, "Maximum a posteriori estimation of room impulse responses," in *ICASSP 2015*, pp. 728–732.

[13] Marco Crocco and Alessio Del Bue, "Room impulse response estimation by iterative weighted l 1-norm," in *2015 23rd European Signal Processing Conference (EUSIPCO)*, 2015, pp. 1895–1899.

[14] Hideaki Kagami, Hirokazu Kameoka, and Masahiro Yukawa, "Joint separation and dereverberation of reverberant mixtures with determined multichannel non-negative matrix factorization," in *ICASSP 2018*, 2018, pp. 31–35.

[15] Deepak Baby, "Supervised speech dereverberation in noisy environments using exemplar-based sparse representations," in *ICASSP 2016*, pp. 156–160.

[16] Jont B Allen and David A Berkley, "Image method for efficiently simulating small-room acoustics," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 943–950, 1979.

[17] Dingzeyu Li, Timothy R Langlois, and Changxi Zheng, "Scene-aware audio for 360 videos," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 111, 2018.

[18] Nicholas J Bryan, "Data augmentation and deep convolutional neural networks for blind room acoustic parameter estimation," *arXiv preprint arXiv:1909.03642*, 2019.

[19] Gavin Kearney, Claire Masterson, Stephen Adams, and Frank Boland, "Dynamic time warping for acoustic response interpolation: Possibilities and limitations," in *2009 17th European Signal Processing Conference*, pp. 705–709.

[20] Akhil Mathur, Anton Isopoussu, Fahim Kawsar, Nadia Berthouze, and Nicholas D Lane, "Mic2mic: using cycle-consistent generative adversarial networks to overcome microphone variability in speech systems," in *Proceedings of the 18th International Conference on Information Processing in Sensor Networks*, 2019, pp. 169–180.

[21] Marco A Martínez Ramírez and Joshua D Reiss, "End-to-end equalization with convolutional neural networks," in *21st International Conference on Digital Audio Effects*, 2018.

[22] Ritwik Giri, Michael L Seltzer, Jasha Droppo, and Dong Yu, "Improving speech recognition in reverberation using a room-aware deep neural network and multi-task learning," in *ICASSP 2015*, pp. 5014–5018.

[23] Gueorgui Pironkov, Stéphane Dupont, and Thierry Dutoit, "I-vector estimation as auxiliary task for multi-task learning based acoustic modeling for automatic speech recognition," in *2016 IEEE Spoken Language Technology Workshop (SLT)*, pp. 1–7.

[24] Suyoun Kim, Bhiksha Raj, and Ian Lane, "Environmental noise embeddings for robust speech recognition," *arXiv preprint arXiv:1601.02553*.

[25] Elad Hoffer and Nir Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*. Springer, 2015, pp. 84–92.

[26] Kaizhi Qian, Yang Zhang, Shiyu Chang, Xuesong Yang, Dinei Florêncio, and Mark Hasegawa-Johnson, "Speech enhancement using bayesian wavenet.," in *Interspeech*, 2017, pp. 2013–2017.

[27] Dario Rethage, Jordi Pons, and Xavier Serra, "A wavenet for speech denoising," in *ICASSP 2018*, 2018, pp. 5069–5073.

[28] Jiaqi Su, Adam Finkelstein, and Zeyu Jin, "Perceptually-motivated environment-specific speech enhancement," in *ICASSP 2019*, pp. 7015–7019.

[29] Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu, "Wavenet: A generative model for raw audio.," in *SSW*, 2016, p. 125.

[30] Michael Buhrmester, Tracy Kwang, and Samuel D Gosling, "Amazon's mechanical turk: A new source of inexpensive, yet high-quality, data?," *Perspectives on psychological science*, vol. 6, no. 1, pp. 3–5, 2011.